Multilingual Semantic Distance:

Automatic Verbal Creativity Assessment in Many Languages

John D. Patterson[1], Hannah M. Merseal[1], Dan R. Johnson[2], Sergio Agnoli[3], Matthijs Baas[4], Brendan S. Baker[1], Baptiste Barbot[5], Mathias Benedek[6], Khatereh Borhani[7], Qunlin Chen[8], Julia F. Christensen[9], Giovanni Emanuele Corazza[10], Boris Forthmann[11], Maciej Karwowski[12], Nastaran Kazemian[7], Ariel Kreisberg-Nitzav[13], Yoed N. Kenett[13], Allison Link[1], Todd Lubart[14], Maxence Mercier[14], Kirill Miroshnik[15], Marcela Ovando-Tellez[16], Ricardo Primi[17], Rogelio Puente-Díaz[18], Sameh Said-Metwaly[19,20], Claire Stevenson[4], Meghedi Vartanian[21,22], Emannuelle Volle[16], Janet G. van Hell[1], & Roger E. Beaty[1]

[1] Pennsylvania State University, University Park, Pennsylvania, United States

[2] Washington & Lee University, Lexington, VA, United States

[3] University of Trieste, Trieste, Italy

[4] University of Amsterdam, Amsterdam, Netherlands

[5] UCLouvain, Ottignies-Louvain-la-Neuve, Belgium

[6] University of Graz, Graz, Austria

[7] Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran

[8] Southwest University, Chongqing, China

[9] Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

[10] University of Bologna, Bologna, Italy

[11] University of Münster, Münster, Germany

[12] University of Wroclaw, Wroclaw, Poland

[13] Technion – Israel Institute of Technology, Haifa, Israel

[14] Université Paris Cité and Univ Gustave Eiffel, LaPEA, Boulognev-Billancourt, France

[15] Saint Petersburg State University, Saint Petersburg, Russia

[16] Sorbonne University, Paris Brain Institute (ICM), INSERM, France

[17] Universidade São Francisco, and EduLab21, Ayrton Senna Institute, São Paulo, Brazil

[18] Universidad Anáhuac México, Estado de México, Mexico

[19] KU Leuven, Leuven, Belgium

[20] Damanhour University, Damanhour, Egypt

[21] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[22] Day Clinic for Cognitive Neurology, University of Leipzig Medical Center, Leipzig, Germany

Author Note

This research was supported in part by the following grants: Austrian Science Fund (FWF; P23914); National Natural Science Foundation of China (31800919); National Council on Scientific and Technological Development (Brazil; 310909/2017-1); São Paulo Research Foundation (Brazil; 2018/10933-8); National Science Centre Poland (UMO-2016/22/E/HS6/00118); Agence Nationale de la Recherche (France; ANR-19-CE37-0001-01); Jacobs Foundation Fellowship (2018 1288 12); German Research Foundation (209933838, CRC 1052/3, Project A1, AV/MS, WI 3342/3-1); Friedrich-Naumann Foundation for Freedom Scholarship; Max Planck Society; Cognitive Sciences & Technologies Council; Abbas Foundation Test Development Funds; ABC Talent Grant; National Science Foundation (United States; DRL-1920653; DUE IUSE-1726811; DUE 1561660; DUE-2155070).

Correspondence should be addressed to Roger E. Beaty, 140 Moore Building, University Park, PA 16802. Email: rebeaty@psu.edu.

Acknowledgements

We are grateful to the many student research assistants who helped with data collection, processing, and scoring, including Victor Altmayer and Raha Golestani.

Abstract

Creativity research commonly involves recruiting human raters to judge the originality of responses to divergent thinking tasks, such as the Alternate Uses Task (AUT). These manual scoring practices have benefitted the field, but they also have limitations, including labor-intensiveness and subjectivity, which can adversely impact the reliability and validity of assessments. To address these challenges, researchers are increasingly employing automatic scoring approaches, such as distributional models of semantic distance. However, semantic distance has primarily been studied in English-speaking samples, with very little research in the many other languages of the world. In a multi-lab study (N = 6,522 participants), we aimed to validate semantic distance on the AUT in 12 languages: Arabic, Chinese, Dutch, English, Farsi, French, German, Hebrew, Italian, Polish, Russian, and Spanish. We gathered AUT responses and human creativity ratings (N = 107,672 responses), as well as criterion measures for validation (e.g., creative achievement). We compared two deep learning-based semantic models—Multilingual Bidirectional Encoder Representations from Transformers (MBERT) and Cross-lingual Language Model RoBERTa (XLMR)—to compute semantic distance and validate this automated metric with human ratings and criterion measures. We found that the top-performing model for each language correlated positively with human creativity ratings, with correlations ranging from medium to large across languages. Regarding criterion validity, semantic distance showed small-to-moderate effect sizes (comparable to human ratings) for openness, creative behavior/achievement, and creative self-concept. We provide open access to our multilingual dataset for future algorithmic development, along with Python code to compute semantic distance in 12 languages.

Keywords: creativity assessment; cross-linguistic analysis; distributional semantic modeling; natural language processing; semantic distance

Multilingual Semantic Distance:

Automatic Verbal Creativity Assessment in Many Languages

When evaluating the originality of ideas on verbal creativity tasks, such as the Alternate Uses Task (AUT)—which prompts participants to produce original and unusual uses for objects—a common method is to ask human raters for their subjective judgments. Subjective creativity scoring, and other methods based on the Consensual Assessment Technique (Benedek et al., 2013; Cseh & Jeffries, 2019; Silvia et al., 2008), have been valuable for the field. But their application often comes at a considerable cost: rating thousands of ideas—as is common in creativity studies—requires a substantial investment of time and effort, which can slow the pace of research (waiting for raters to complete the arduous task of rating) and adversely impact reliability and validity of test scores, particularly when raters disagree or provide unreliable ratings (e.g., due to fatigue; Forthmann et al., 2017; Rönkkö & Cho, 2020). Moreover, the necessity of multiple human raters limits the applicability of creativity assessments in education, where human resources are scarce.

To address these issues, researchers are increasingly exploring computational methods for automating the scoring process (Acar et al., 2021; Acar & Runco, 2014; Beaty et al., 2021; Beaty & Johnson, 2021; Bendetowicz et al., 2018; Bossomaier et al., 2009; Dumas, Organisciak, et al., 2020; Forster & Dunbar, 2009; Forthmann & Doebler, 2022; Gray et al., 2019; Heinen & Johnson, 2018; Johnson, Kaufman, et al., 2021; Olson et al., 2021; Paulus et al., 1970; Prabhakaran et al., 2014; Rafner et al., 2022; Shute & Rahimi, 2021; Stevenson et al., 2020; Sung et al., 2022; Volle, 2018; Yu et al., 2022; Zedelius et al., 2019). One promising approach employs distributional semantic models to compute *semantic distance*, which quantifies how “far away” an idea is from common ideas (Kenett, 2019). In English samples, semantic distance correlates positively with human creativity ratings and other measures of creativity (Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Prabhakaran et al., 2014; Stevenson et al., 2020), highlighting its construct validity and utility to the field.

Despite its promise, the application of semantic distance in creativity research has been largely restricted to English-speaking research participants (cf. Bendetowicz et al., 2018; Forthmann et al., 2018; C. Liu et al., 2021; Stevenson et al., 2020; Sung et al., 2022). Very little psychometric research has been conducted using semantic distance in the many other languages of the world. This English-only bias limits the accessibility of powerful automatic scoring approaches—which, by extension, slows the pace of research in non-English speaking countries—and reduces the comparability and transparency of research findings across languages. In the present project, we aimed to address this issue by forming an international consortium of researchers who conduct research on creative thinking. Each lab contributed responses to the widely-used AUT in their respective language, along with human creativity ratings and criterion measures (e.g., creative achievement), representing data from 12 languages collected in 12 countries, with over 6,000 participants and over 100,000 AUT responses. Using state-of-the-art multilingual semantic models, we aimed to validate semantic distance for creativity assessment beyond English.

Human and Machine Assessment of Verbal Creativity

When conducting research on creative thinking, researchers must decide on a method for evaluating the many responses that participants produce on idea generation tasks, such as the AUT, within a fixed amount of time (Acar & Runco, 2019). Depending on the sample size, AUT studies can yield hundreds or thousands of responses, which then need to be scored in some way before they can be analyzed for the purpose of the study. A straightforward way of scoring AUT responses is to simply count them (i.e., fluency): participants who have many ideas receive a high fluency score, and those who have fewer ideas receive a low fluency score. Yet fluency alone cannot speak to the quality of ideas: participants who produced many unoriginal ideas (e.g., common uses for objects on the AUT) would still receive a high fluency score, raising questions about the construct validity of the task (Benedek et al., 2013; Forthmann et al., 2020; Nusbaum et al., 2014). To assess the quality of ideas, researchers can calculate flexibility

(i.e., the number of semantic categories visited) and/or uniqueness (i.e., the statistical infrequency of a response). Yet these metrics have been criticized for their strong dependence on fluency (flexibility scales with fluency; more categories, more ideas) and sample size (uniqueness decreases with larger samples—a rare instance of adverse impact from large sample size; Forthmann, Paek, Dumas, Barbot, & Holling, 2019).

An alternative scoring approach that overcomes these issues is the subjective scoring method (Silvia et al., 2008). Subjective scoring is based on the Consensual Assessment Technique (CAT), an approach that relies on "experts" to provide their personal evaluation, often with minimal guidance on what constitutes a creative idea or product (Amabile, 1983; Cseh & Jeffries, 2019). According to the CAT, the extent to which raters independently agree is critical to determining the reliability of a creativity assessment (Hennessey, 1994). When applied to the AUT, the CAT often involves asking raters to judge the originality of responses, e.g., using a 1 (*not at all creative*) to 5 (*highly creative*) scale. A large literature has demonstrated the reliability and validity of subjective scoring on the AUT and other creative thinking tasks (e.g., Benedek et al., 2013; Jauk, Benedek, & Neubauer, 2014; Silvia et al., 2008), highlighting the psychometric strengths of subjective scoring.

In addition to its strengths, however, subjective scoring has some limitations. Perhaps the most notable limitation is the labor cost of conducting research: scoring hundreds or thousands of ideas is quite costly in terms of time and human resources. Creativity researchers often rely on undergraduate research assistants—which requires recruiting, training, and retaining a team of raters to provide careful and consistent subjective ratings (Benedek et al., 2013)—thus constraining the pace of research by the availability of qualified raters. Importantly, such volunteers are not always accessible in university research labs, and they are rarely available in other educational settings (e.g., primary schools), preventing educators from efficiently testing creativity in their classrooms. In addition, the process of rating thousands of responses can lead to rater fatigue, adversely impacting the reliability of ratings (Forthmann et

al., 2017). There is also the issue of rater disagreement: raters do not always agree on what they find creative (Ceh et al., 2022), reflecting a source of noise that violates a central tenet of the CAT, i.e., that a creativity assessment produces reliable and valid scores to the extent that experts agree (Amabile, 1983; Cseh & Jeffries, 2019). Moreover, when different labs use different scoring procedures, this limits the comparability and transparency of research findings, and may also impact replicability.

To address the challenges of subjective scoring, a growing number of researchers are exploring automated scoring methods (Acar et al., 2021; Acar & Runco, 2014; Beaty & Johnson, 2021; Dumas et al., 2021; Dumas, Organisciak, et al., 2020; Dumas & Runco, 2018; Forthmann & Doebler, 2022; Gray et al., 2019; Johnson, Kaufman, et al., 2021; Kenett, 2019; Olson et al., 2021; Paulus et al., 1970; Prabhakaran et al., 2014; Rafner et al., 2022; Shute & Rahimi, 2021; Stevenson et al., 2020; Sung et al., 2022; Yu et al., 2022; Zedelius et al., 2019). One prominent approach is to compute semantic distance using distributional semantic models—a class of natural language processing tools that quantifies conceptual similarity in texts (Günther et al., 2019; Jackson et al., 2022). Semantic distance reflects “how far” two concepts are from each other in a high dimensional semantic space by computing the cosine similarity between concepts, reflecting their co-occurrence in large collections of natural language. Thus, if two concepts co-occur frequently (e.g., *coffee—drink*), they have a low semantic distance (.46); likewise, if two concepts co-occur infrequently (e.g., *coffee—write*), they have a high semantic distance (.93)[1]. The application of semantic distance in creativity assessment aligns with the associative theory of creativity, i.e., the view that creative thinking requires connecting distantly associated concepts, and that creative people have highly connected memory structures that facilitate remote conceptual combination (Kenett, 2019).

[1] Semantic distance values were computed by SemDis (semdis.wlu.psu.edu) using the GloVe model.

Semantic distance has received psychometric support for producing reliable and valid scores in English-speaking samples, with several studies reporting positive correlations between semantic distance scores and human creativity ratings obtained on creative thinking tasks (Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Yu et al., 2022). Early studies on semantic distance used latent semantic analysis (LSA)—a "count" model that computes semantic distance by counting the number of co-occurrences of word pairs (Bossomaier et al., 2009; Forster & Dunbar, 2009). For example, Prabhakaran et al. (2014) applied semantic distance to a word association task and found that participants who generated more semantically distant word associations (when instructed to "be creative") tended to perform better on other tests of creative thinking and report higher levels of creative achievements, as well as higher levels of openness to experience, demonstrating the construct validity of semantic distance scores using LSA. Semantic distance of responses on an analogical reasoning task was also found to modulate activity in left frontopolar cortex (an area implicated in analogical reasoning; Green et al., 2012). Finally, semantic distance is well-correlated with idea originality, distinct from idea fluency (Dumas & Dunbar, 2014).

Recently, Beaty and Johnson (2021) extended this work by incorporating multiple semantic models into the computation of semantic distance. In addition to LSA, the authors explored "predict" models (i.e., neural networks that predict missing words from surrounding context words), aiming to improve the generalizability of semantic distance for creativity assessment by capturing a more diverse range of semantic models and text corpora, instead of LSA alone (Kenett, 2019). In five studies (three studies using the AUT and two studies with word association tasks), Beaty and Johnson (2021) found consistently large correlations between semantic distance scores and human ratings of creativity and novelty, as well as measures of creative performance and personality. Other studies have reported similar findings, such as Dumas et al. (2020), who found high correlations between AUT semantic distance and human ratings; and Dumas, Doherty, and Organisciak (2020), who found that a group of

creative professionals (actors) produced more semantically distant AUT responses than a less creative control group.

Additional text mining methods have been developed to assess other aspects of creative performance, such as elaboration (Dumas et al., 2021), originality (Acar & Runco, 2014), and flexibility (Johnson, Cuthbert, et al., 2021), as well as free association on the forward flow task (Beaty et al., 2021; Gray et al., 2019) and narrative creativity on creative writing tasks (Johnson, Kaufman, et al., 2021; Zedelius et al., 2019). Several open access tools have been released to improve the accessibility of these automated methods, including web applications for scoring the AUT and other verbal creativity tasks (https://openscoring.du.edu; https://semdis.wlu.psu.edu); the free association task, "forward flow" (http://www.forwardflow.org); and the divergent association task (DAT; https://www.datcreativity.com).

The Present Research

Semantic distance is a promising alternative to subjective creativity scoring, with increasing evidence to support its reliability and validity, and a growing number of open-access resources for researchers to facilitate automated assessment. To our knowledge, however, semantic distance-based creativity assessment has focused almost entirely on English-speaking participants, with very little psychometric work in the many other languages of the world. This disparity constitutes a major barrier to accessibility and diversity in the field, slowing the pace of research in non-English speaking countries who are subject to the limits and bottlenecks of subjective scoring. Moreover, the acceleration of semantic distance research in English-speaking countries—in the absence of parallel progress in other languages—is problematic from a comparative perspective: any conclusions based on English-speaking samples (derived from semantic models) will not necessarily generalize to other languages (derived from subjective scoring).

In the present research, we sought to address this issue by validating semantic distance in many different languages. To this end, we formed a global consortium of creativity researchers working in 12 different languages in 15 different countries. These 12 languages entail 6 different language families: Germanic (Dutch, English, German), Romance (French, Italian, Spanish), Slavic (Polish, Russian), Semitic (Arabic, Hebrew), Indo-Iranian (Farsi), and Sinitic (Chinese). Each lab contributed previously collected data from the AUT, as well as human creativity ratings, and in most cases, additional measures for validation purposes (e.g., creative achievement). Our collective dataset includes data from over 6,500 participants, with over 107,000 AUT responses.

We tested the efficacy of two multilingual semantic models—established by the machine learning literature (Conneau et al., 2020)—for computing semantic distance: Multilingual BERT (MBERT) and Cross-lingual Language Model RoBERTa (XLMR). Both models are multilayer transformer neural networks. The key innovation introduced by the transformer architecture is a set of attentional mechanisms that allow the model to differentially weigh words in a sentence and adapt its word vector representations based on the surrounding word context (Vaswani et al., 2017). This enables the model to represent words in a nuanced, context-dependent way and handle cases of polysemy (e.g., *dish* as something you cook vs. as something you put away in a cupboard). Importantly, both MBERT and XLMR were pretrained on at least 100 different languages, including all 12 languages assayed in the current work.

For each model, and for each AUT response, we compute the maximum associative distance (MAD), i.e., the most semantically distant word in a response (Yu et al., 2022). Typically, participants use multiple words to describe their ideas on the AUT. Like other methods for computing semantic distance, MAD computes the semantic distance between the AUT object (e.g., *rope*) and all words in the response. However, whereas other methods combine all of the semantic distance values into one (i.e., compositional vectors; e.g., multiplicative and additive), MAD retains only the most semantically distant word in the

response, removing the rest. Yu et al. (2022) examined the reliability and validity of scores obtained with the MAD method in English. Across three studies, MAD significantly outperformed current state-of-the-art compositional methods in predicting human creativity ratings and criterion measures (e.g., openness, creative achievement). Here, we apply the MAD method to our multilingual dataset, testing the extent to which person-level MAD scores correlate with human ratings and criterion measures across languages.

Method

The materials, anonymized data, and code from this project are available on OSF (https://osf.io/5cy9n/?view_only=36f893c28bcc4ceb8404913bb9471aeb).

Participants

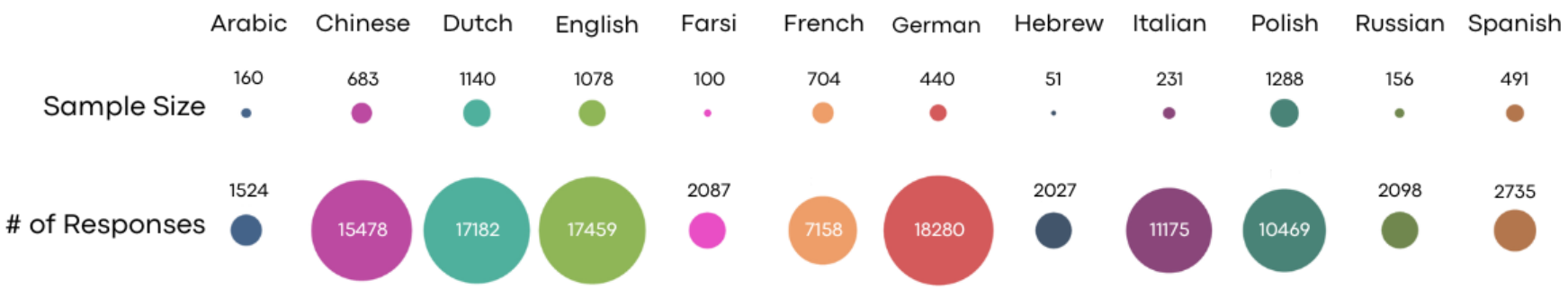
The current study is part of an international project that aims to develop automated tools for verbal creativity assessment. Initially, the last author emailed researchers from various countries to invite them to contribute data to the project, with the goal of collecting data from as many languages as possible. The invitation requested AUT responses, subjective creativity ratings, and validation measures (e.g., openness to experience, creative achievement). In addition to inviting researchers via email, a call to contribute data was made at the 7th annual meeting of the Society for the Neuroscience of Creativity. We received 30 datasets, with a combined sample size of 6,522, reflecting data from 22 labs and 12 languages: Arabic, Chinese, Dutch, English, Farsi, French German, Hebrew, Italian, Polish, Russian, and Spanish (see Figure 1). Several datasets came from published studies, whereas others have not been used for publication.

Procedure

Participants completed various cognitive tests and self-report scales across the 30 datasets. Some studies were completed online, and others were completed in-person. All participants completed a version of the AUT; most participants also completed additional measures, which were used to validate semantic distance scores in the present study.

Figure 1

Sample size and number of AUT responses for each language



The present study focused on three common validation measures in the creativity literature: openness to experience, creative behavior/achievement, and creative self-efficacy. Table 1 indicates which datasets had these measures; not all datasets had all variables available, and different measures were used across datasets (e.g., openness scales). Additional measures not analyzed in the present study are available on OSF.

Alternate Uses Task (AUT). The AUT is a widely used measure of creative thinking. Participants are presented with objects and asked to think of uses for them. Several different items (e.g., *brick, rope*) were used across the datasets. Task duration also varied considerably, with a median duration of 2.5 minutes (range: 8 seconds to 10 minutes). Task and rater instructions, task durations, and items included in each of the 30 datasets are reported in the Supplementary Materials on OSF. A majority of the studies instructed participants to "be creative", which has been shown to improve the reliability and validity of the AUT (Acar et al., 2020; Said-Metwaly et al., 2020).

Across all datasets, AUT responses were scored using the subjective scoring method (Benedek et al., 2013; Silvia et al., 2008). Raters received different scoring guidance across the 30 datasets, though a majority used some variation of publicly available scoring guidelines (https://osf.io/vie7s/), which emphasize uncommonness, remoteness, and cleverness. They rated the quality of AUT responses using a Likert scale, which varied across studies (e.g., 1 = *not at all creative*, 5 = *very creative*; see Supplemental Materials for rater instructions). The number of raters also varied across datasets (median = 3 raters, range = 1-45)[2]. For each dataset, creativity ratings were first z-scored (for each rater; to account for rater severity), then z-scored ratings were averaged across raters for each response; these response level z-scored ratings were then averaged at the item level (e.g., *box, rope*), and averaged again at the participant level. Table 1 lists (for each dataset) the number of raters and their reliability

[2] The Hebrew responses were rated across a pool of 45 raters such that each response was rated by 8 judges.

(Intraclass Correlation Coefficient, with 95% confidence interval, via the 'irrNA' package in R; Brueckl & Heuer, 2021). To demonstrate the internal consistency of the AUT items within a given language (in cases where more than one item was used in a given language, and participants completed multiple items), within-subject Pearson correlations between items with respect to human creativity ratings are available on the OSF page ('item-item_correlations'). While there is a wide range of item-item correlations, the correlations generally land within the .2-.4 range.

Creative behavior. Self-reported creative activities and achievements were assessed with various scales across datasets. The Creative Achievement Questionnaire (Carson et al., 2005) assesses creative accomplishments across ten domains. The Inventory of Creative Activities and Achievements (Diedrich et al., 2018) measures both hobbies and accomplishments in eight domains. The Creative Activity and Accomplishment Checklists (Okuda et al., 2016) assesses activities and achievements in six domains.

Creative self-concept. The Short Scale of Creative Self (Karwowski, 2014) was used to assess two components of creative self-concept: creative self-efficacy (CSE; 6 items; e.g., "I am good at proposing original solutions to problems") and creative personal identity (CPI; 5 items; e.g., "Being a creative person is important to me"). Participants respond to a series of questions using a 5-point Likert scale (1 = *definitely not*, 5 = *definitely yes*). For each subscale, a total score is derived by averaging the items.

Openness to experience. The Big 5 trait openness to experience was assessed using several different scales, including the Big Five Aspects Scale (DeYoung et al., 2007), Big Five Inventory (John et al., 1991), NEO PI-R (McCrae et al., 2005), and Ten-Item Personality Inventory (Gosling et al., 2003).

Table 1

Validation Measures, Raters, and Creativity Rating ICC by Dataset

Dataset	Measures	Raters	ICC
Arabic1	CAQ, openness	1	N/A [N/A, N/A]
Chinese1	N/A	4	0.49 [0.48, 0.51]
Chinese2	CAQ, openness	4	0.64 [0.6, 0.67]
Dutch1	N/A	2	0.81 [0.8, 0.82]
Dutch2	N/A	2	0.94 [0.94, 0.95]
Dutch3	CAQ	2	0.87 [0.86, 0.89]
Dutch4	CAQ	3	0.85 [0.84, 0.86]
English1	ICAA.act, ICAA.ach, CAQ, openness, CPI, CSE	4	0.84 [0.83, 0.85]
English2	ICAA.act, openness	4	0.77 [0.76, 0.78]
English3	ICAA.act, ICAA.ach, openness, CPI, CSE	3	0.56 [0.53, 0.58]
English4	CAQ, openness	3	0.72 [0.7, 0.74]
English5	openness	3	0.78 [0.76, 0.79]
English6	openness	3	0.64 [0.62, 0.66]
Farsi1	N/A	3	0.69 [0.66, 0.71]
Farsi2	N/A	3	0.75 [0.71, 0.77]
French1	ICAA.act, ICAA.ach	4	0.8 [0.78, 0.81]
French2	N/A	3	0.64 [0.58, 0.7]
French3	CSE, CPI, openness	3	0.75 [0.73, 0.76]
French4	CSE, CPI, openness	3	0.8 [0.78, 0.81]
German1	CAQ, ICAA.act, ICAA.ach, openness	4	0.71 [0.7, 0.72]
German2	N/A	3	0.78 [0.76, 0.79]
German3	N/A	3	0.86 [0.86, 0.87]
Hebrew1	N/A	45	0.88 [0.87, 0.89]
Italian1	CAAC, openness	2	0.89 [0.89, 0.9]
Italian2	CAAC, openness	2	0.88 [0.87, 0.89]
Polish1	CPI, CSE, CAQ, openness	3	0.82 [0.81, 0.83]
Polish2	CPI, CSE, CAQ, openness	2	0.6 [0.55, 0.63]
Russian1	N/A	3	0.72 [0.7, 0.75]
Russian2	openness	3	0.79 [0.75, 0.83]
Spanish1	CPI, CSE	3	0.74 [0.73, 0.76]

Note. CAAC = Creative Activity and Accomplishment Checklists; CAQ = Creative Achievement Questionnaire; CPI = Creative Personal Identify; CSE = creative self-efficacy; ICAA = Inventory of Creative Activities and Achievements (act = activities; ach = achievements); N/A = not available. ICC values were computed using a two-way random effects model with average rater consistency via the 'irrNA' package in R (Brueckl & Heuer, 2021).

Automated Assessment

Semantic Distance Computation. AUT responses are commonly expressed with multiple words, requiring a methodological decision of how to compute semantic distance and aggregate corresponding word vectors (e.g., multiplication or addition; Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020). Recently, Yu and colleagues tested an alternative approach to composition—MAD (maximum associative distance)—which computes semantic distance between the prompt word (e.g., *box*) and all words in a response (e.g., *cut the box into circular coasters for drinks*). Critically, unlike compositional methods, MAD only retains the most semantically-distant word in the response (i.e., *coasters;* semantic distance = .99)[3], removing all other words for the final semantic distance score. Yu et al. (2022) found that MAD significantly outperformed the compositional approach (multiplying word vectors) in predicting human ratings and criterion measures in English samples (e.g., openness, creative achievement). We thus employ the MAD approach to compute semantic distance at the person level (as opposed to item or response level) in the present study.

Semantic Models. The pretrained models used in the present work were obtained via the 'HuggingFace Transformers' suite of the 'PyTorch' package for the Python programming language. Within the 'HuggingFace Transformers' suite, the variant of MBERT we used was 'bert-base-multilingual-cased'; for XLMR we used the 'xlm-roberta-large' variant. Although the models share many similarities—they are both multilayer bidirectional encoder transformers, are both trained with a fill-mask objective (i.e., some words in every training sentence are masked and its goal is to fill in the blanks), and are both shown to perform well on cross-lingual tasks

[3] Semantic distance values were computed by SemDis (semdis.wlu.psu.edu) using the GloVe model.

(Conneau et al., 2020; Wu & Dredze, 2019)—the two models were primarily chosen for their differences.

For one, the models differ in their capacity and depth. While MBERT has twelve 768-dimension layers and a total of 110 million parameters, XLMR consists of 24 layers of 1027 dimensions and a total of 355 million parameters. Thus, XLMR is twice as deep and has over three times as much capacity (parameters) as MBERT. Aside from being larger, XLMR is also trained on a different dataset. MBERT is trained on the 104 languages with the largest Wikipedia databases while XLMR is trained on cleaned CommonCrawl data that covers 100 languages. CommonCrawl is an archive, of steadily-increasing size, that is produced by an internet bot which systematically explores webpages; cleaning the CommonCrawl archive is thought to increase the influence of 'low-resource' languages (those with less web presence) on the model's knowledge of those languages by orders of magnitude (Wenzek et al., 2020). Third, while both models are trained to fill in a missing 15% of each training text input, there are two primary differences in how the models were trained: (1) MBERT consistently masks the same words across training presentations while XLMR dynamically changes which 15% of the words are masked across presentations of the same text; and (2) MBERT has an added training objective that XLMR does not—it binds two sentence-level representations together and has to decide if they were next to each other in the source text or not. Researchers that introduced the monolingual precursor to XLMR (i.e., RoBERTa; Liu et al., 2019) found this 'next sentence prediction' training task did not aid model performance and subsequently dropped it from the training regime in the RoBERTa framework. Given XLMR is built on this framework, the next sentence prediction objective is omitted from the training of XLMR as well.

As noted above, there are several differences between the backends of the two semantic models we explore in the current work. Naturally, this precludes systematic comparison between the two models. Our aim, however, is not systematic comparison. Instead, our aim is to test the extent to which currently available cutting-edge multilingual transformer

models can predict human-evaluated creativity across a diversity of languages. As such, in principle, diversity in model size, training regime, and source of training data enhances the probability that the assayed languages will be represented effectively, particularly considering the language-specific model fitting procedure noted below.

Model Fitting. As we cannot know *a priori* which of the two models will best predict creativity for a given language, conducting a computational experiment to find the best-fitting model for each language is key. An added modelling complication comes from the fact that each model has *n* layers, each of which consists of a *d* dimensional vector of activations. These layer activations are used to compute semantic distance (i.e., MAD) for the model. Recent work (Johnson, Kaufman, et al., 2021) using BERT large—a 24-layer English monolingual transformer model similar to those in this study—found that layers 6 and 7 provided the best fit to human creativity ratings for narratives. However, the best single layer, or layer pair, for predicting AUT creativity is unknown in this novel multilingual case using new semantic models.

In the present work, we perform a computational experiment with the aim of maximizing creativity-rating prediction for each language by searching over two factors: (1) model type (MBERT vs. XLMR) and (2) layer preference (best layer vs. top-2 averaged). To ensure that the best-fitting models generalize as well as possible to new AUT responses within each language, we use *k*-fold cross-validation to select the best model (i.e., which model type and layer preference is best for a given language). For each language, we split the data into fifths (i.e., $k = 5$) along the participant level such that data from a fifth of the participants was in each split. Model and layer preference selection proceeded independently. Determination of each model's layer preference proceeded first and consisted of two steps: (1) ascertaining which two layers yielded MAD values that correlated highest with human ratings and (2) comparing whether a best or top-2 averaged approach provided a better fit to human ratings. For the first step, each model went through five 'selection' iterations. In each iteration, the model was provided data from 4/5ths of the participants; correlations between the MADs of each of the *L* model layers (12

for MBERT, 24 for XLMR) and human creativity ratings for AUT responses were stored. This process was repeated another four times, where the held-out fifth of the participants was unique each time. The selection iterations resulted in five folds of *L* correlations from each model.

The two layers that had the highest average correlation with human ratings, across the five folds, were then used for the second step. The MAD scores from the best layer and the top two layers (averaged) were pitted against one another on each fifth of the data (i.e., the held-out 'test' sets). Whichever approach had the highest average correlation across the five held-out test sets won and was deemed the layer preference of the model. Finally, to determine the best fit overall, the five (*k*) test set correlations from the winning layer preference of each model were Fisher Z transformed and averaged. Whichever layer-preference-filtered model (MBERT or XLMR) had a higher average correlation was selected as the best-fitting model.

Results

Best-Fitting Model Settings

The best-fitting models for each language are shown in Table 2. As can be seen, there was substantial diversity between languages in terms of the semantic model, and layer preference for MAD computation, that best accounted for human creativity ratings. Exactly half (6/12) of the languages were best fit by the smaller MBERT. However, excluding English, a slight majority (6/11) of the languages were best fit by XLMR.

With respect to the best approach for computing MAD values—either using only the layer that correlated best with human creativity ratings or averaging MAD values from the top two layers—the layer preference was also diverse, though biased in favor of the top-two approach, suggesting multiple layers often contained information important for predicting human ratings. Eight of the 12 languages (67%) were best fit by the top2 solution, while only four languages (33%) were fit best by computing MAD values based on the single most performant layer. The breakdown is comparable when excluding English (64% were best fit by the top2 approach, 36% were best fit by the single best layer).

Table 2

Best-fitting Models by Language

Language	Model	Layer.Preference	Correlation
Arabic	XLMR	Best	0.28
Chinese	MBERT	Top2	0.24
Dutch	MBERT	Best	0.47
English	MBERT	Top2	0.52
French	MBERT	Top2	0.24
Farsi	XLMR	Top2	0.23
German	XLMR	Best	0.41
Hebrew	XLMR	Top2	0.23
Italian	MBERT	Top2	0.37
Polish	XLMR	Best	0.35
Russian	MBERT	Top2	0.38
Spanish	XLMR	Top2	0.40

*Note. '*Correlation' reflects the mean Pearson correlation across all *k* folds for the best-fitting model (where *k* correlations were Fisher Z transformed, averaged, then back transformed to a Pearson correlation). 'XLMR' = Cross-lingual Language Model RoBERTa. 'MBERT' = Multilingual BERT. 'Best' indicates MAD scores were derived solely from the layer that correlated highest with human ratings; 'Top2' indicates MAD scores were derived by averaging MAD scores across the two layers with the highest correlations to the rating data.

Considering the conjunction of semantic model *and* layer preference, in cases where XLMR was the best-fitting model, 50% of the languages were best fit by a single layer preference while the remaining 50% of languages were best fit by a top2 preference. In contrast, MBERT was more biased toward using the top2 approach. Five out of the six cases where MBERT best captured human creativity ratings employed the top2 approach; only in one case was the single-layer approach preferred.

Correlations to Human Creativity Ratings

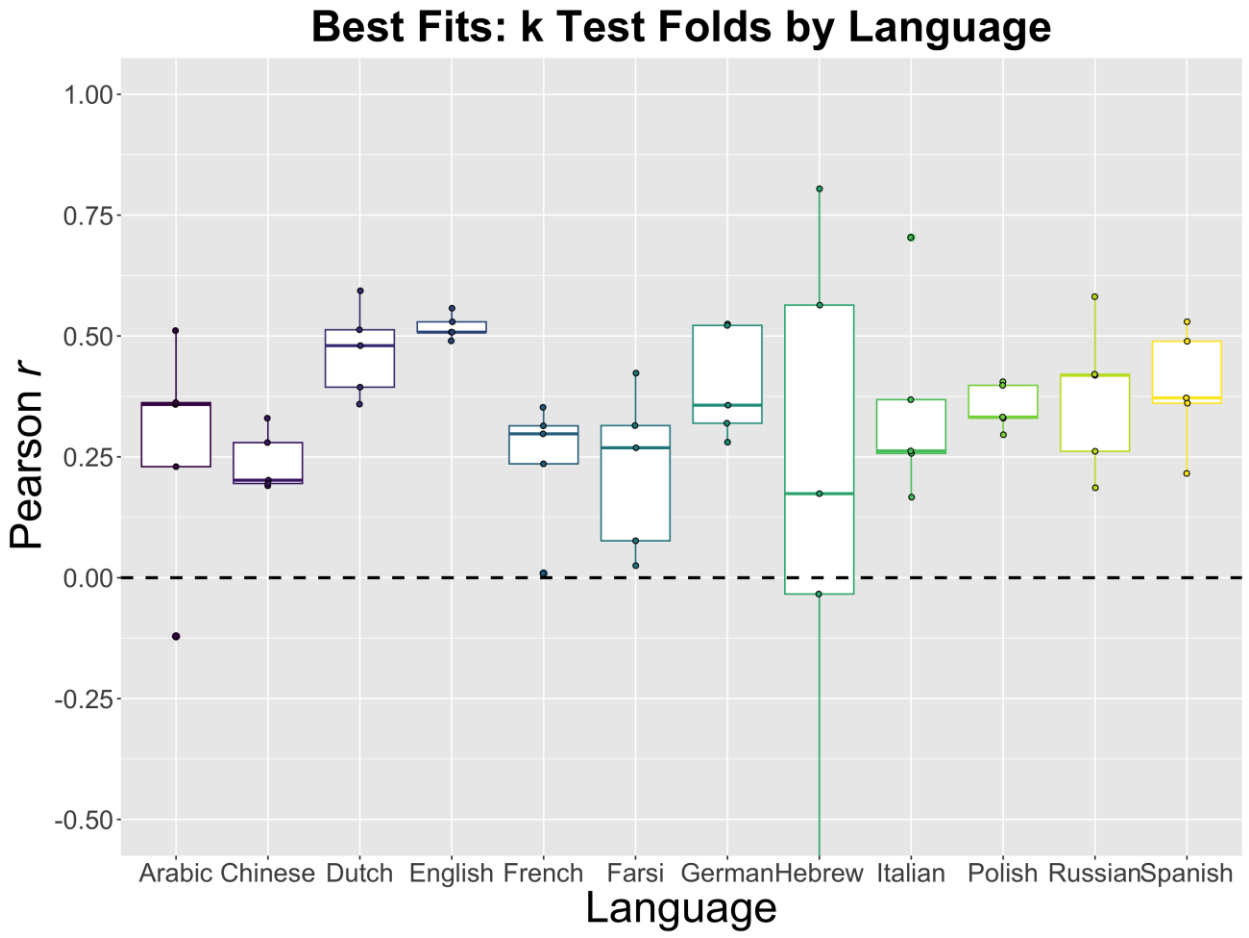
Person-level correlations between the best-fitting model for each language and human creativity ratings for each of the *k* test data subsets are depicted in Figure 2 (median

correlations are represented by the darkened horizontal lines in the boxplots). Additionally, the mean correlation, across test folds, for each language can be viewed in Table 2.

As can be seen, the directionality of the observed correlation coefficients was generally positive at both the average and individual test fold levels. The positive directionality observed at the average level did not appear to be driven by outliers, as the median test fold correlation for each language was also positive.

Figure 2

Pearson Correlations Across All k Test Folds for Each Language



Note. The x-axis shows the 12 languages; the y-axis shows the Pearson correlation with human ratings for the 5 folds of each dataset. The dots shown for each language represent the correlation between MAD values and human ratings for each of the *k* test splits; the darkened line in the boxplot shows the median correlation value. Note: the correlation for the Hebrew dataset (*k*-fold) that is not shown is r = -.62 (n = ~10 for each fold in the Hebrew dataset).

Importantly, across most languages, the semantic distance-human rating correlations were medium to large in magnitude. Seven out of the 12 languages displayed moderate to strong correlations between model-predicted and human-provided ratings (i.e., mean $r \geq .30$). The largest model-human correlation was observed for English ($r = .52$). Dutch achieved the largest non-English correlation ($r = .47$) and was followed closely by German and Spanish ($r = .41$, $r = .40$, respectively). Smaller correlations ($r < .30$) were observed for Arabic, Farsi, French, Chinese, and Hebrew.

Surprisingly, an English bias did not emerge. Looking at the average correlations (Table 2), the magnitude difference between English and the largest non-English coefficient was minimal ($r_{\text{English-Dutch}} = .05$). For perspective, the difference in magnitude between the two highest non-English correlations was roughly the same ($r_{\text{Dutch-German}} = .06$).

Criterion Validity

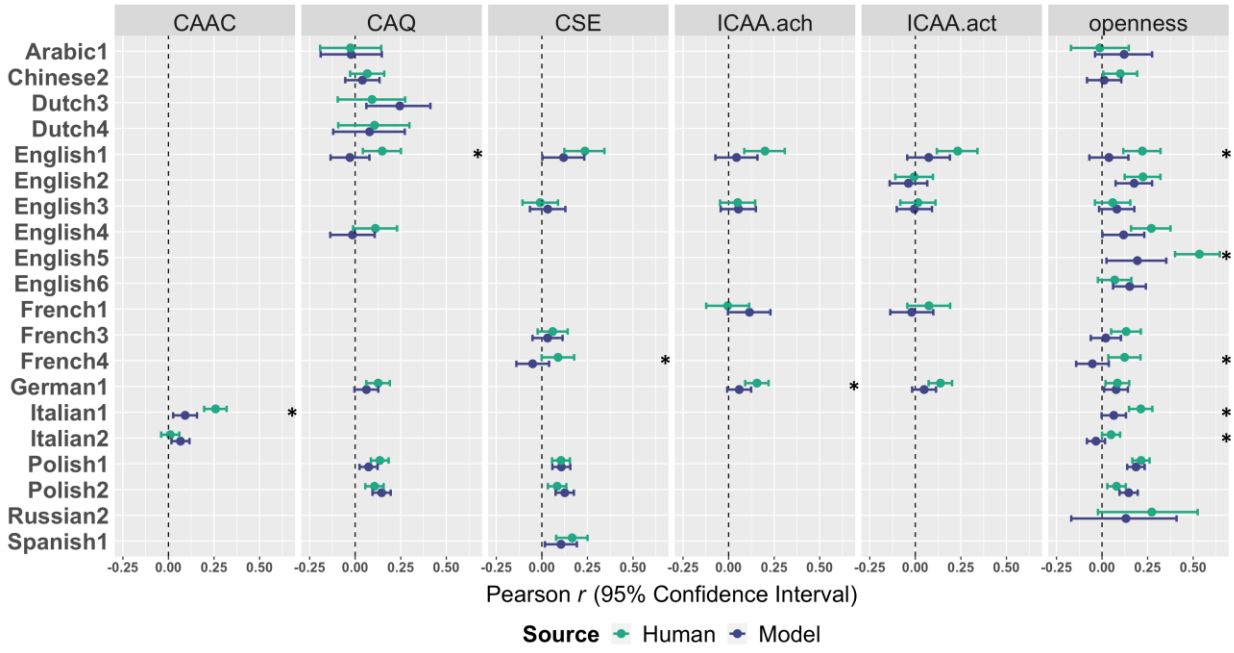
Next, we assessed criterion validity of the top-performing semantic models/layers (for each language) with respect to their correlations with three commonly-used measures of creative behavior and personality: openness to experience, creative behavior/achievement, and creative self-concept. We also computed correlations between the three criterion measures and human creativity ratings as a baseline comparison for the semantic models. Figure 3 displays the Pearson correlations (and their 95% confidence intervals) for semantic distance, human ratings, and the three validation measures across datasets/languages.

Regarding openness, correlations were modest but largely comparable to human ratings across languages—the human and semantic-distance correlations did not significantly differ across seven of 11 cases where human correlations significantly exceeded zero. Small and near-zero correlations were found for some datasets (e.g., Chinese, French, Italian), including the English1 dataset, consistent with prior work reporting variable openness—semantic distance correlations in English samples (Beaty, Johnson, et al., 2022; Beaty & Johnson, 2021). Interestingly, German and Polish showed larger criterion validity correlations for openness than

some of the English datasets, indicating that non-English models can occasionally exceed English criterion validity performance (although various between-

Figure 3

Correlations Between Best-fitting Model for Each Dataset and Criterion Measures



Note. Each panel represents a different measure. CAAC = Creative Activity and Accomplishment Checklists; CAQ = Creative Achievement Questionnaire; CPI = Creative Personal Identify; CSE = creative self-efficacy; ICAA = Inventory of Creative Activities and Achievements (act = activities; ach = achievements). Each dot represents the correlation between scores on a given measure and human creativity ratings (green) or MAD values from the best-fitting model (blue) for each dataset. Note that only 111 participants (out of 297 participants) had associated criterion measures in the Dutch3 dataset; the correlations shown reflect only that subset. Asterisks indicate the difference between model and human correlations is significant, though not the degree of significance (i.e., $p < .05$ and $p < .001$ are both represented by a single asterisk). Statistical comparisons between model and human correlations were obtained via the *cocor* package for R (Diedenhofen & Musch, 2015).

sample/language differences could also influence these results, such as the reliability of different personality scales; see *Discussion*).

Regarding creative behavior and achievement, similar trends emerged: correlations were generally small but similar in magnitude to human ratings across languages and measures

(i.e., CAAC, CAQ, and ICAA). The largest semantic distance-creative achievement correlation was found in the Dutch3 sample for the CAQ. Interestingly, the magnitude of this correlation (r = .24) exceeded the magnitude of the human correlation (r = .09), which was also the case in the French1 dataset for ICAA (though not significantly). In only three out of 15 cases did the human correlation significantly exceed the semantic distance correlation for creative achievement (i.e., Italian1, English1, and German1); otherwise, human and model did not differ, suggesting comparable predictive validity for both automated and subjective scoring methods for creative behavior/achievement.

Regarding creative self-concept, we found significant correlations with semantic distance in four of the seven datasets (English1, Polish1 and 2, and Spanish1). These correlations were small and not significantly different from human ratings. Of the three other datasets with available creative self-concept data (French and English), only one showed a significant correlation with human ratings (French4). Thus, criterion validity evidence for semantic distance and creative self-concept, like openness and creative behavior/achievement, largely mirrored the validity evidence for human ratings on the AUT.

Discussion

Creativity research has historically required human raters to manually evaluate the quality of ideas produced on creative thinking tasks (Acar & Runco, 2019). To address the subjectivity and labor cost of manual scoring, machine learning methods have been developed and psychometrically validated—yet such tools have been largely limited to English-speaking people. In the present project, we aimed to expand access to automated creativity assessments beyond English, applying two state-of-the-art multilingual models of semantic distance to the widely used AUT. Across 11 non-English languages (comprising 6 different language families), we found that semantic distance correlated positively with human originality ratings, with variable performance across languages and multilingual semantic models (and their corresponding layer preferences). To further validate multilingual semantic distance, we

examined the correlations between the best-performing model/layers with measures of creative personality and behavior (openness, creative achievement, and creative self-concept), finding generally comparable correlations to human originality ratings. Our results extend automatic creativity assessment beyond the English language, providing a means to accelerate the pace of creativity research and, critically, to facilitate cross-cultural comparisons in verbal creativity.

We conducted a computational experiment to identify the optimal semantic model (XLMR or BERT) and layer combination for each language. Our results yielded an even split between XLMR and BERT, and a mix between the best and top2 layer preference, in terms of the strongest correlations with human ratings for each language. English showed the highest mean correlation with human ratings (across the 5 *k*-folds; $r = .52$). Importantly, the multilingual models also yielded moderate to large correlations with human ratings for several languages, with most effect sizes within the range of .35 to .45 at the person level. The magnitude of these zero-order correlations is consistent with effect sizes reported in previous English samples (Beaty, Johnson, et al., 2022; Beaty & Johnson, 2021), albeit not for all languages. Notably, there was wide variability across languages/datasets on factors that could influence the model's correlation with human ratings, such as sample size and inter-rater agreement. Nevertheless, our results provide a novel demonstration that semantic distance can capture variance in human creativity ratings across languages, opening the door for its application in future studies with non-English samples.

We also sought to validate semantic distance against three common criterion measures: openness to experience, creative behavior/achievement, and creative self-concept. Overall, the correlations between semantic distance and the three criterion measures were modest across languages, ranging from 0 to ~.3, depending on the criterion measure. Notably, although correlations were slightly larger for human ratings on average, the magnitude of semantic distance correlations was generally comparable to correlations with human ratings, and even occasionally exceeded human correlations. The effect sizes are also consistent with English

studies (Beaty, Johnson, et al., 2022; Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Yu et al., 2022), particularly for self-report measures (e.g., creative self-concept), which have found inconsistent correlations with both semantic distance and human ratings on the AUT.

These mixed validity findings for semantic distance may speak more to the limits of the AUT—and its ability to predict real-world creative performance (Stevenson et al., 2021)—than the semantic models themselves, given the comparable correlations found for human ratings and semantic distance. Other verbal tasks, such as creative writing and even simple word association tasks, lend themselves well to semantic distance analysis, and have shown encouraging validity evidence (Beaty et al., 2021; Bendetowicz et al., 2018; Gray et al., 2019; Johnson, Kaufman, et al., 2021; C. Liu et al., 2021; Prabhakaran et al., 2014). The semantic models tested in the present study could be extended to other verbal tasks in several languages, following psychometric evaluation to determine their reliability and validity.

Strengths, Limitations, and Future Directions

The present study is the first cross-cultural validation of semantic distance for verbal creativity assessment. Beyond psychometric validation, our project offers insight into factors that contribute to human evaluations of verbal creativity across languages and cultures. Specifically, we show that semantic distance—an objective indicator of originality based on distributional semantic models—correlates with human ratings of creativity in 11 different non-English languages, indicating that originality may be universally valued across cultures, at least when evaluating ideas on verbal creativity tasks. In addition to the predictive accuracy, a particular strength of this automated scoring approach is the speed with which creativity scores are returned. To illustrate, this automated approach scores responses in approximately half a second, which means this approach can score three AUT items, with five responses each, from 100 participants in 12 minutes (on our hardware).

Our study was bolstered by an international collaboration of creativity researchers who contributed their data to the project, yielding a large and unprecedented collection of verbal

creativity responses in 12 total languages (along with human creativity ratings, validation measures, and experimental materials). In addition to validating semantic distance, we also provide a large-scale, cross-cultural validation of common criterion measures (e.g., openness to experience) with respect to human creativity ratings on the AUT. To facilitate future research on automated creativity assessment, we provide open access to the data and materials on OSF, along with the machine learning models used to produce semantic distance scores in 12 languages.

Our study has a few limitations worth noting. First, the 30 datasets varied along several dimensions that could impact the results, including sample size, number of raters, instructions given to raters/participants, AUT items and task duration, and validation measures. Regarding sample size, we used a *k*-folds approach, which split datasets into five equal folds, to find the optimal model/layer combination for each language. Thus, languages with larger samples (e.g., Dutch) tended to show less variance in the *k*-folds analysis than languages with smaller samples (e.g., Hebrew). The Hebrew dataset (N = 51 participants) was particularly affected by the *k*-folds analysis (N = 10 per *k*-fold), which yielded a large/negative correlation for one fold. We thus urge caution when interpreting the results of the Hebrew dataset. Likewise, with respect to the number of raters, the Arabic dataset had a single rater, which may have skewed the model/layer selection process.

Regarding instructions and task duration, prior work has demonstrated that both factors play important roles in divergent thinking assessment, particularly instructions to “be creative” (Acar et al., 2020; Said-Metwaly et al., 2020). Although most datasets used some version of “be creative” instructions (see Supplemental Materials), there were still notable differences across languages that may have played a role. Regarding AUT items, recent evidence indicates that different AUT items yield different semantic distance values (Beaty, Johnson, et al., 2022), as well as different fluency values and human originality ratings (Beaty, Kenett, et al., 2022; Forthmann et al., 2016). Moreover, the method used to aggregate across multiple AUT

responses (e.g., averaging scores within-person) plays an important role in explaining validity correlations (i.e., fluency confound; Beaty, Johnson, et al., 2022; Benedek et al., 2013; Silvia et al., 2008). Taken together, these differences between datasets make comparative analysis challenging, but a systematic analysis of how such moderators (e.g., task duration or instructions) relate to the alignment of model predictions with human creativity ratings represents a promising direction for future work.

Another limitation of the modeling approach in the present work is that it may not be sensitive to response appropriateness. The models in this study were designed to predict human creativity ratings but not necessarily response appropriateness. It is reasonable to suspect that human raters would deem inappropriate/incoherent responses as less creative, but the modeling approach in this work would most likely not. Given the semantic distance method applied to the model's latent word activations, it is likely that an incoherent, random string of words would yield very high MAD scores. Future work should seek to quantify the extent to which the current approach is vulnerable to such 'gaming' of the system and construct safeguards if necessary.

A potential limitation may also be found in the usage of MAD as the semantic distance metric used in our modeling approach. While extant research indicates that MAD outperforms other approaches to computing semantic distance (Yu et al., 2022 [preprint]), it is also true that MAD discards a lot of information by retaining only the largest distance found between the AUT prompt item and the words of the response. Additional work comparing the performance of different semantic distance metrics across languages will help optimize cross-lingual performance in future multilingual modeling endeavors.

We encourage future research to extend this work by conducting comparative analysis of automated creativity assessments. Subsequent studies could address the limitations of the current project—which was constrained by the availability of existing data—by controlling study parameters across languages as much as possible (e.g., items, task/rater instructions). The field

of comparative linguistics is rich with theories and methods for exploring how people use language differently across cultures, and powerful tools from natural language processing are now available to study a range of psychological processes in text, from emotion to creativity (Jackson et al., 2022). Although researchers have been studying cross-cultural differences in creativity for some time, e.g., the relative importance of novelty and appropriateness in Eastern vs. Western cultures (Ivancovsky et al., 2018; Niu & Sternberg, 2002), we look forward to more work along these lines in other cultures and languages that are less well-represented in the creativity literature. A related issue concerns measurement invariance across cultures: assessments of creative potential, like the AUT, may not necessarily measure creative potential the same across cultures (Guo et al., 2021). Thus, researchers should carefully consider the equivalence of creativity assessments before making inferences about cross-cultural differences.

The current work may also serve as an important foundation for further forays into automated multilingual assessment. A key strength of the present work is that it releases a curated, multilingual anthology of over 107k human-rated AUT responses (with validation measures) to the research community (OSF: https://osf.io/5cy9n/?view_only=36f893c28bcc4ceb8404913bb9471aeb). Given its size and accessibility, the dataset holds the potential to serve as a yardstick against which novel computational approaches to multilingual creativity assessment can be compared (i.e., a 'benchmark' dataset).

There are several promising targets for novel computational approaches to multilingual creativity prediction. One target stems from the rapid pace of innovation in machine learning. Though we deployed models that were state-of-the-art among those publicly available, new models of promise are not far around the corner (e.g., GPT-3; Stevenson, Smal, Baas, Grasman, & van der Maas, 2022). Second, semantic distance in the current work was based on the pretrained versions of the selected models; the models' word representations were derived

only from what they learned in the fill-mask training task. However, multiple studies show that the representations of pretrained transformers are adaptable, and can be tuned to different cross-lingual tasks (Conneau et al., 2020; see also Organisciak et al., 2022 [pre-print] for finetuning on English AUT responses). A promising next step is thus to train models specifically on the task of multilingual creativity prediction. Moreover, future studies should explore compositional approaches to aggregating word vectors (e.g., additive and multiplicative models), as has been done in previous work (Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020), testing whether the MAD approach—which shows higher validity evidence in English (Yu et al., 2022)—is similarly optimal for other languages. Last, the current work investigated one- and two-layer approaches for computing semantic distance—motivated by prior work (Johnson, Kaufman, et al., 2021). However, exploring different many-layer protocols for computing semantic distance will be important in subsequent lines of inquiry.

Finally, future studies should also expand the scope of automated assessments to creativity tasks beyond the AUT, such as narrative creativity (Fletcher & Benveniste, 2022)—particularly given the recent success of text analysis tools applied to short stories in English-speaking samples (Johnson, Kaufman, et al., 2021; Toubia et al., 2021; Zedelius et al., 2019)—with an eye toward diversifying creativity research and increasing the accessibility of creativity assessments beyond English.

References

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C. “Tedd,” & Organisciak, P. (2021). Applying Automated Originality Scoring to the Verbal Form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly*. https://doi.org/10.1177/00169862211061874

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

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. https://doi.org/10.1037/aca0000231

Acar, S., Runco, M. A., & Park, H. (2020). What Should People Be Told When They Take a Divergent Thinking Test? A Meta-Analytic Review of Explicit Instructions for Divergent Thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 39–49. https://doi.org/10.1037/aca0000256

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, *45*(2), 357–376. https://doi.org/10.1037/0022-3514.45.2.357

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., Kenett, Y. N., Hass, R. W., & Schacter, D. L. (2022). Semantic memory and creativity: the costs and benefits of semantic memory structure in generating original ideas. *Https://Doi.Org/10.1080/13546783.2022.2076742*, 1–35. https://doi.org/10.1080/13546783.2022.2076742

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

Bendetowicz, D., Urbanski, M., Garcin, B., Foulon, C., Levy, R., Bréchemier, M. L., Rosso, C., De Schotten, M. T., & Volle, E. (2018). Two critical brain networks for generation and combination of remote associations. *Brain*, *141*(1), 217–233. https://doi.org/10.1093/brain/awx294

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

Bossomaier, T., Harre, M., Knittel, A., & Snyder, A. (2009). A semantic network approach to the Creativity Quotient (CQ). *Creativity Research Journal*, *21*(1), 64–71. https://doi.org/10.1080/10400410802633517

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? *Journal of Creative Behavior*, *56*(1), 41–54. https://doi.org/10.1002/jocb.515

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *1*, 346–350. https://doi.org/10.48550/arxiv.2006.13979

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

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between Facets and Domains: 10 Aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

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., Doherty, M., & Organisciak, P. (2020). The psychology of professional and student actors: Creativity, personality, and motivation. *PLOS ONE*, *15*(10), e0240728. https://doi.org/10.1371/JOURNAL.PONE.0240728

Dumas, D., Organisciak, P., & Doherty, P. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*.

Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four Text-Mining Methods for

Measuring Elaboration. *The 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

Dumas, D., & Dunbar, K. N. (2014). Understanding Fluency and Originality: A latent variable perspective. *Thinking Skills and Creativity, 14,* 56–67. https://doi.org/10.1016/j.tsc.2014.09.003

Fletcher, A., & Benveniste, M. (2022). A new method for training creativity: narrative as an alternative to divergent thinking. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/NYAS.14763

Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the cognitive science society* (pp. 602–607). Cognitive Science Society.

Forthmann, B., & Doebler, P. (2022). *Fifty Years Later and Still Working: Rediscovering Paulus et al.'s (1970) Automated Scoring of Divergent Thinking Tests*. https://doi.org/10.31234/OSF.IO/BYJ8C

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32. https://doi.org/10.1016/j.intell.2016.03.005

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. (2018). Application of latent semantic analysis to divergent thinking is biased by elaboration. *Journal of Creative Behavior*, *53*(4), 559–575. https://doi.org/10.1002/jocb.240

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2019). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*. https://onlinelibrary.wiley.com/doi/abs/10.1111/bjep.12325

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), 683–699. https://doi.org/10.1111/bjep.12325

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

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. *American Psychologist*, *74*(5), 539–554. https://doi.org/10.1037/amp0000391

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264–272. https://doi.org/10.1037/a0025764

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Guo, Y., Lin, S., Guo, J., Lu, Z. (Laura), & Shangguan, C. (2021). Cross-cultural measurement invariance of divergent thinking measures. *Thinking Skills and Creativity*, *41*, 100852. https://doi.org/10.1016/J.TSC.2021.100852

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. (1994). The Consensual Assessment Technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, *7*(2), 193–208. https://doi.org/10.1080/10400419409534524

Ivancovsky, T., Kleinmintz, O., Lee, J., Kurman, J., & Shamay-Tsoory, S. G. (2018). The neural underpinnings of cross-cultural differences in creativity. *Human Brain Mapping*, *39*(11), 4493–4508. https://doi.org/10.1002/HBM.24288

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

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), 95–105. https://doi.org/10.1002/per.1941

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—versions 4a and 54*.

Johnson, D. R., Cuthbert, A. S., & Tynan, M. E. (2021). The neglect of idea diversity in creative idea generation and evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 125–135. https://doi.org/10.1037/aca0000235

Johnson, D. R., Kaufman, J. C., Baker, B. S., Barbot, B., Green, A. E., Hell, J. van, Kennedy, E.,

Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2021). Extracting Creativity from Narratives using Distributional Semantic Modeling. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/FMWGY

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

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

Liu, C., Ren, Z., Zhuang, K., He, L., Yan, T., Zeng, R., & Qiu, J. (2021). Semantic association ability mediates the relationship between brain structure and human creativity. *Neuropsychologia*, *151*, 107722. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2020.107722

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. https://doi.org/10.48550/ARXIV.1907.11692

McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment*, *84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

Niu, W., & Sternberg, R. (2002). Contemporary Studies on the Concept of Creativity: the East and the West. *The Journal of Creative Behavior*, *36*(4), 269–288. https://doi.org/10.1002/J.2162-6057.2002.TB01069.X

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

Okuda, S. M., Runco, M. A., & Berger, D. E. (2016). Creativity and the Finding and Solving of Real-World Problems: *Http://Dx.Doi.Org/10.1177/073428299100900104*, *9*(1), 45–53. https://doi.org/10.1177/073428299100900104

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25). https://doi.org/10.1073/PNAS.2022340118/SUPPL_FILE/PNAS.2022340118.SAPP.PDF

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2022). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. [Preprint]. http://dx.doi.org/10.13140/RG.2.2.32393.31840

Paulus, D. H., Renzulli, J. S., & Archambault, F. X. (1970). *Computer Simulation of Human Ratings of Creativity. Final Report.* (No. 9-A-032). https://files.eric.ed.gov/fulltext/ED060658.pdf

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

Rafner, J., Biskjær, M. M., Zana, B., Langsford, S., Bergenholtz, C., Rahimi, S., Carugati, A., Noy, L., & Sherson, J. (2022). Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities. *Creativity Research Journal*, *34*(1), 28–54. https://doi.org/10.1080/10400419.2021.1971447

Rönkkö, M., & Cho, E. (2020). An Updated Guideline for Assessing Discriminant Validity: *Https://Doi.Org/10.1177/1094428120968614*, *25*(1), 6–14. https://doi.org/10.1177/1094428120968614

Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., & Van den Noortgate, W. (2020). Testing Conditions and Creative Performance: Meta- Analyses of the Impact of Time Limits and Instructions. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 15–38. https://doi.org/10.1037/ACA0000244

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

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

Stevenson, C., Baas, M., & van der Maas, H. (2021). A Minimal Theory of Creative Ability. *Journal of Intelligence*, *9*(1), 1–19. https://doi.org/10.3390/JINTELLIGENCE9010009

Stevenson, C., Smal, I., Baas, M., Dahrendorf, M., Grasman, R., Tanis, C., Scheurs, E., Sleiffer, D., & Maas, H. van der. (2020). Automated AUT scoring using a Big Data variant of the Consensual Assessment Technique. In *Final technical report for the Abbas Foundation Test Development Funds 2016* (Vol. 11). https://dare.uva.nl/search?identifier=13ad004a-1b61-45a0-8a9a-56d7a165d7ef

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's Creativity to the (Alternative Uses) Test*. arXiv. https://doi.org/10.48550/ARXIV.2206.08932

Sung, Y.-T., Cheng, H.-H., Tseng, H.-C., Chang, K.-E., & Lin, S.-Y. (2022). Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/ACA0000450

Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts

their success. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(26), 2021. https://doi.org/10.1073/PNAS.2011695118/SUPPL_FILE/PNAS.2011695118.SAPP.PDF

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009. https://doi.org/10.48550/arxiv.1706.03762

Volle, E. (2018). Associative and controlled cognition in divergent thinking: Theoretical, experimental, neuroimaging evidence, and new directions. In *The Cambridge Handbook of the Neuroscience of Creativity* (pp. 333–360). Cambridge University Press. https://doi.org/10.1017/9781316556238.020

Wenzek, G., Lachaux, M. A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 4003–4012. https://doi.org/10.48550/arxiv.1911.00359

Wu, S., & Dredze, M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 833–844. https://doi.org/10.48550/arxiv.1904.09077

Yu, Y., Beaty, R., Forthmann, B., Beeman, M., Cruz, J. H., & Johnson, D. R. (2022). A mad method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *PsyArXiv*. https://doi.org/10.31234/OSF.IO/VGXPK

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