Author/s:
Olson, JA;Nahas, J;Chmoulevitch, D;Cropper, SJ;Webb, ME

Title:
Naming unrelated words predicts creativity

Date:
2021-06-22

Citation:
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.

Persistent Link:
https://hdl.handle.net/11343/281146

Naming unrelated words predicts creativity

Jay A. Olson[a,1], Johnny Nahas[b], Denis Chmoulevitch[b], Simon J. Cropper[c], and Margaret E. Webb[c]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138; [b]Department of Psychology, McGill University, Montreal, QC, Canada H3A 1G1; and [c]Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC 3010, Australia

Edited by Paul Verhaeghen, Georgia Institute of Technology, Atlanta, GA, and accepted by Editorial Board Member Randall W. Engle April 9, 2021 (received for review October 26, 2020)

Several theories posit that creative people are able to generate more divergent ideas. If this is correct, simply naming unrelated words and then measuring the semantic distance between them could serve as an objective measure of divergent thinking. To test this hypothesis, we asked 8,914 participants to name 10 words that are as different from each other as possible. A computational algorithm then estimated the average semantic distance between the words; related words (e.g., cat and dog) have shorter distances than unrelated ones (e.g., cat and thimble). We predicted that people producing greater semantic distances would also score higher on traditional creativity measures. In Study 1, we found moderate to strong correlations between semantic distance and two widely used creativity measures (the Alternative Uses Task and the Bridge-the-Associative-Gap Task). In Study 2, with participants from 98 countries, semantic distances varied only slightly by basic demographic variables. There was also a positive correlation between semantic distance and performance on a range of problems known to predict creativity. Overall, semantic distance correlated at least as strongly with established creativity measures as those measures did with each other. Naming unrelated words in what we call the Divergent Association Task can thus serve as a brief, reliable, and objective measure of divergent thinking.

creativity | divergent thinking | semantic distance | computational scoring

Think of three words that are as different from each other as possible. Choosing these words relies on generating remote associations while inhibiting common ones, according to two dominant theories of creativity (1, 2). Associative theories posit that creative people have a semantic memory structure that makes it easier to link remote elements (3–6). Executive theories focus on top-down control of attention; creative solutions arise from monitoring and inhibiting common associations (2, 7). Based on these theories, we hypothesized that the simple act of naming unrelated words may reliably measure verbal creativity.

Creativity has two main psychological components, convergent thinking and divergent thinking, which work together when generating creative output. Convergent thinking tasks measure the ability to assess several stimuli and arrive at the most appropriate response, such as the optimal solution to a problem (3, 8–10). These tasks tend to be easier to score since there is a small set of correct answers. In contrast, divergent thinking tasks typically use open-ended questions that measure one's ability to generate various solutions (11–13). They usually require longer text-based responses and are thus harder to objectively score. The most common divergent thinking measure is the Alternative Uses Task (14, 15), in which participants generate uses for common objects such as a paper clip or a shoe. Using a common method of scoring (16), raters then judge the responses based on three components:

- flexibility, the number of distinct categories of uses generated;
- originality, how rare each use is relative to the rest of the sample, which is particularly important for creativity (17, 18); and
- fluency, how many uses are generated in total.

Perhaps fittingly, there are diverse ways to score tests of divergent thinking (11, 19–24). Many of the manual scoring methods, however, have several drawbacks. The scoring is laborious and time intensive (11), and multiple judges are required to assess reliability, which adds to the effort (25). Further, the scoring is sample dependent (23, 25); originality is scored in a relative, nonabsolute manner. Thus, a participant's responses will be more or less rare (and more or less original) depending on the other responses in the sample. Finally, the scoring does not account for cultural differences; uses of objects vary in commonality across cultures and at different times. The use of a paper clip to change the smart card in a smartphone, for example, is now one of the most common responses, although it was rare a decade ago. Ratings of originality will thus vary by country and year. This issue makes it difficult to accurately judge responses from multicultural or international samples or to assess how divergent thinking changes over time.

To address these limitations, recent efforts have moved toward using computational algorithms to score task responses (11, 17, 26, 27). Compared with manual scoring, computational methods may also clarify the theoretical grounding of the measures since the assumptions required to score the responses must be made explicit in the program code (11, 13). Researchers have successfully automated the scoring of a broad range of creativity measures that examine noun–verb pairs (28), synonyms (29), and chains of word associations (12, 30). Many of these computational methods can generate scores similar to human ratings, including on the Alternative Uses Task (11, 17, 26, 27).

The multiword answers of the Alternative Uses Task are well suited for assessing appropriateness and usefulness, which are important aspects of creativity (31), but they may be less suited for computational scoring. Such scoring methods typically involve computing the semantic distance between each of the words independent of their order (e.g., using Latent Semantic

Significance

Many traditional measures of creativity require time-intensive and subjective scoring procedures. Their scores are relative to the specific sample, which makes multicultural or international assessments difficult. Our results show that a shorter and simpler task with automatic and objective scoring may be at least as reliable at measuring verbal creativity. This finding enables assessments across larger and more diverse samples with less bias.

Author contributions: J.A.O., J.N., D.C., S.J.C., and M.E.W. designed research; S.J.C. and M.E.W. performed research; J.A.O. contributed new reagents/analytic tools; J.A.O., J.N., and D.C. analyzed data; and J.A.O., J.N., D.C., S.J.C., and M.E.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. P.V. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: jay.olson@mail.mcgill.ca.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2022340118/-/DCSupplemental.

Published June 17, 2021.

Analysis) (32). Common methods also remove stop words (e.g., the, which, of) in order to increase the consistency of the scoring (27, 33). These practices can obscure meaning, which is a well-known issue in the field of artificial intelligence. Word meaning often depends on context and order, as demonstrated by the sentence "Time flies like an arrow; fruit flies like a banana" (34). In the Alternative Uses Task, using a pen to "record a break" is less original than using one to "break a record." Additionally, the number of words used to describe a concept can also influence semantic distance (33). Using a magazine to "do a collage" vs. to "cut out pieces to glue to a poster for a collage" will result in different scores. Thus, a task that uses single-word responses and instructs participants to consider all meanings of the words may reduce these problems and result in more reliable scoring.

Accordingly, we aimed to develop a measure of divergent thinking that focused on remote associations using single-word responses. Our goal was to develop a task that is brief; is easy to implement; and offers objective, automatic, and absolute scoring.

The Divergent Association Task

Our proposed measure, the Divergent Association Task (DAT), asks participants to generate 10 nouns that are as different from each other as possible in all meanings and uses of the words. We then compute the semantic distances between them; words that are used in similar contexts have smaller distances. The words cat and dog, for example, would be close to each other since they are often used together, whereas cat and thimble would not. We computed the semantic distance using an algorithm called GloVe (35), which has previously been used to score the Alternative Uses Task (26, 27). We used a freely available model that was pretrained on the Common Crawl corpus, which contains the text of billions of web pages (35).

To provide some redundancy (as described below), we keep only the first seven valid words that participants provide. The DAT score is the transformed average of the semantic distances between these words. In particular, we compute the semantic distance (i.e., cosine distance) between all 21 possible pairs of the seven words, take the average, and then multiply it by 100. The full algorithm code is available online (https://osf.io/bm5fd/). The minimum score (zero) occurs when there is no distance between the words: that is, when all of the words are the same. The theoretical maximum score (200) would occur when the words are as different from each other as possible. In practice, scores commonly range from 65 to 90 and almost never exceed 100. Scores under 50 are often due to misunderstanding the instructions, such as naming opposites (e.g., day and night) rather than unrelated words. In this way, the score can be intuitively thought of as a grade on an examination; under 50 is poor, the average is between 75 and 80, and 95 is a very high score. Fig. 1 shows example words and their corresponding scores.

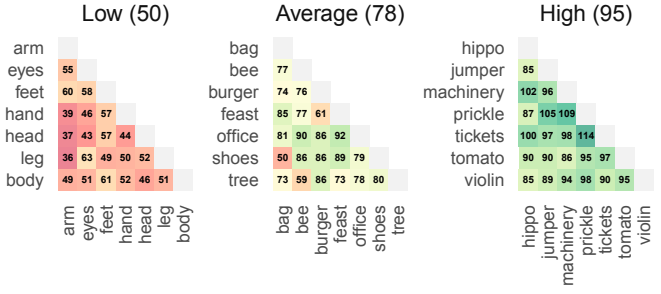

Fig. 1. Examples of participant responses and their corresponding DAT scores. The score is the transformed average of the semantic distances between each pair of words.

This operationalization of divergent thinking is grounded in associative and executive control theories of creativity. Higher scores would demonstrate a greater ability to draw upon more remote associations (3–5) or to inhibit overly related associations (2, 7). In Study 1, we tested this hypothesis by comparing the DAT with two other measures of creativity: the Alternative Uses Task (15) and the Bridge-the-Associative-Gap Task (36). In Study 2, we tested how these scores vary by demographics and whether they correlate with other measures related to divergent thinking in a larger dataset (9, 37). These studies assessed whether semantic distance could be a reliable indicator of divergent thinking.

Results

Number of Words. The DAT asks participants to name 10 unrelated words, but we only required a subset of these to provide a buffer for mistakes. This way, if participants mistyped a few words or chose some that were absent from the model, we could still compute an overall score. In Study 1A, all participants ended up providing at least seven valid words (Fig. 2*A*), so we used this number in all of the samples to compute the DAT score; additional words were discarded.

Using these first seven words, the average DAT score for Study 1A was 78.38 (SD = 6.35), which was similar to the subsequent samples (*SI Appendix*, Table S1). Most participants finished the DAT in a minute and a half, with a median response time of 88.21 s (SD = 66.46). The task is thus shorter than many traditional measures of divergent thinking.

Correlations with Other Creativity Measures. In Study 1A, using a manually screened subset of the data in which participants followed the instructions most closely (*Materials and Methods*), the DAT correlated well with the Alternative Uses Task (Fig. 2*B*). In particular, the DAT correlated with flexibility [$r(55) = 0.51$ $[0.29, 0.68]$, $P < 0.001$] and originality [$r(55) = 0.50$ $[0.28, 0.68]$, $P < 0.001$], but we did not see the same correlation with fluency [$r(55) = 0.21$ $[-0.05, 0.45]$, $P = 0.057$]. Using the full dataset with no manual screening, we saw positive correlations across all three measures (flexibility: $r = 0.34$ $[0.18, 0.48]$; originality: $r = 0.32$ $[0.16, 0.46]$; fluency: $r = 0.22$ $[0.06, 0.37]$).

Participants also completed the Bridge-the-Associative-Gap Task, a test of convergent thinking in which participants see two words (e.g., giraffe and scarf) and need to find a third one that relates to both (e.g., neck). Raters then judged the appropriateness of the provided words. We saw a positive correlation between the DAT and appropriateness in the manually screened subsample [$r(54) = 0.34$ $[0.08, 0.55]$, $P = 0.006$] as well as in the full dataset [$r(136) = 0.22$ $[0.06, 0.38]$, $P = 0.004$].

Study 1B attempted to replicate these findings in another dataset, without any manual screening. We again saw positive correlations with the Alternative Uses Task [flexibility: $r(223) = 0.35$ $[0.23, 0.46]$, $P < 0.001$; originality: $r(223) = 0.32$ $[0.20, 0.43]$, $P < 0.001$; fluency: $r(223) = 0.30$ $[0.17, 0.41]$, $P < 0.001$] and appropriateness ratings in the Bridge-the-Associative-Gap Task [$r(203) = 0.23$ $[0.10, 0.36]$, $P < 0.001$].

To assess test–retest reliability, in Study 1C participants completed the DAT during laboratory visits 2 wk apart for an unrelated study (38). Test–retest reliability was high [$r(48) = 0.73$ $[0.57, 0.84]$, $P < 0.001$]; this reliability resembled that of completing the same Alternative Uses Task items 1 mo later, as scored by raters (r = 0.61 to 0.70) or an algorithm (r = 0.49 to 0.80) (39).

In our preregistered Study 2, due to time constraints, we used a shortened version of the Alternative Uses Task. Participants were asked to name a single "new and imaginative use" for a common object across two trials. A confirmatory test showed a positive correlation between the DAT and manually scored

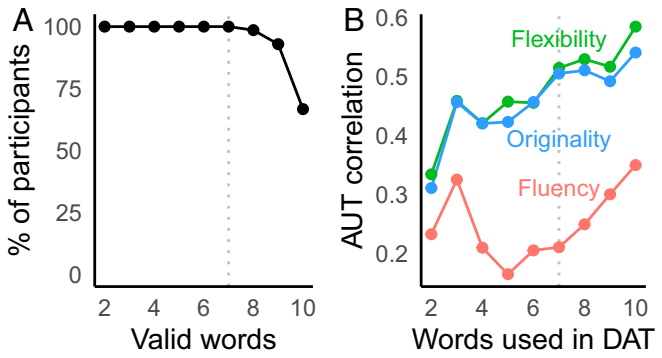

Fig. 2. (*A*) Percentage of participants included when requiring different numbers of valid words in the DAT score and (*B*) corresponding correlations with the Alternative Uses Task (AUT) scores for each number of words. For example, the first point on each graph shows the (*A*) inclusion rate and (*B*) correlation when using only the first two valid words provided. Using the first 7 of 10 words balanced high correlations with a high inclusion rate.

originality [$r(353) = 0.13$ $[0.03, 0.23]$, $P = 0.006$]. The magnitude was lower than in Study 1 likely because the shortened version of the task had less precision.

Demographics. In Study 2, confirmatory tests also showed that the DAT scores differed only slightly by basic demographic variables such as age and gender (*SI Appendix*, Table S2). Scores were slightly higher in females and peaked in their twenties (Fig. 3). All demographic factors combined explained 1% of the total variation in the model, suggesting that the DAT varies little by these basic demographics.

Problems. Participants in Study 2 also completed various problems known to predict creativity: two items of a convergent thinking task (the Compound Remote Associates Test), one insight problem, and one analytical problem. On average, participants answered 2.11 (SD = 1.03) of the 4 items correctly. An exploratory test showed that those who correctly completed more items had higher DAT scores [$r(348) = 0.16$ $[0.05, 0.26]$, $P = 0.003$].

Enjoyment. After each measure of Study 2, participants reported how much pleasure they experienced on a zero to five scale. Participants enjoyed the DAT the most, with an average rating of 3.56 [3.52, 3.59] compared with the Alternative Uses Task items ($M = 2.65$ $[2.62, 2.67]$) or any of the other problems ($M = 2.74$ $[2.63, 2.84]$) (*SI Appendix*, Fig. S1). These results match reports from our pilot testing; participants often described the task as enjoyable or fun (at least for a cognitive task).

Comparison with Other Correlations. Across Studies 1 and 2, we saw positive correlations between the DAT and the other creativity measures. These sample correlations were generally at least as strong as the correlations among the other established measures (Fig. 4).

Discussion

Our results suggest that simply asking participants to name unrelated words can serve as a reliable measure of divergent thinking. We compared performance on this task with established creativity measures—the Alternative Uses Task and the Bridge-the-Associative-Gap Task—as well as related measures of insight and analytical problem solving. The correlations between the DAT and these measures tended to be at least as high as the correlations among the other established measures themselves, demonstrating strong convergent validity. The highest correlations were between the DAT and the Alternative Uses Task (Studies 1A and 1B). Test–retest reliability was also high over a span of 2 wk (Study 1C). Overall, the evidence supports semantic distance as a reliable measure of divergent thinking. Although the precise mechanism underlying this link is unclear, the DAT may indirectly measure the extent or efficiency of the association network, as suggested by associative (3, 4, 6) and executive control theories (2, 20).

Performance on the DAT varied little by demographic measures (Study 2). Other studies have shown mixed results relating demographics and creativity. For example, studies are inconclusive about the impact of gender differences (40, 41). Abraham (42) reports highly mixed findings, with approximately half of studies reporting no differences in creativity by gender and the other half reporting mixed results, with possibly higher scores in women. There are also inconsistent findings regarding age; some studies indicate that performance on creativity tasks increases as domain-specific knowledge and vocabulary increase (43), while others find the opposite (44). Generally, however, studies find that creativity declines in late adulthood (41). In our sample, the low variation by demographic variables suggests that the DAT can be used across ages and genders without modification.

Strengths. The DAT resolves a number of limitations in the creativity literature. Compared with manually scored tasks, the DAT scoring is automatic and objective, allowing researchers to collect large samples with little effort and no rater bias. Further, the scoring is absolute and not sample dependent, making it conducive to comparing diverse populations. The DAT may also reduce some biases seen in other tests of divergent thinking. Experiential bias occurs when one's past experience influences the diversity of responses (45). When listing uses of a brick in the Alternative Uses Task, for example, a brick layer may provide different responses than a lawyer, leading to more uncontrollable variation. Similarly, different object prompts in the Alternative Uses Task can lead to different responses with varying reliability between computational scoring and manually scored responses (27). The DAT avoids these issues by giving an open-ended prompt and using a model trained on an international corpus (i.e., global website data). Relatedly, fluency contamination occurs when high fluency can artificially inflate originality scores; the more responses that participants list, the more likely some of them will be unique (22). The DAT avoids this issue by requiring all participants to generate the same number of responses.

Finally, the task is short; most participants completed it in a minute and a half and rated it as more enjoyable than every other creativity task used in our study. The task can also provide instant

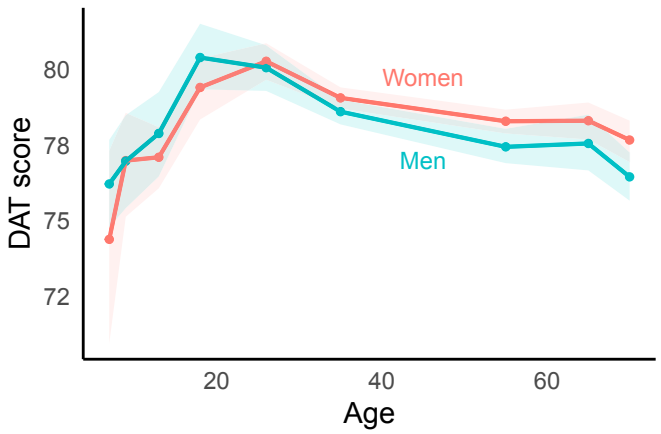

Fig. 3. DAT scores by age and gender. Scores peaked in young adulthood, and females showed slightly higher overall scores than males. Dots show means, and bands show 95% CIs. Age was approximated by the minimum value of each bin.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

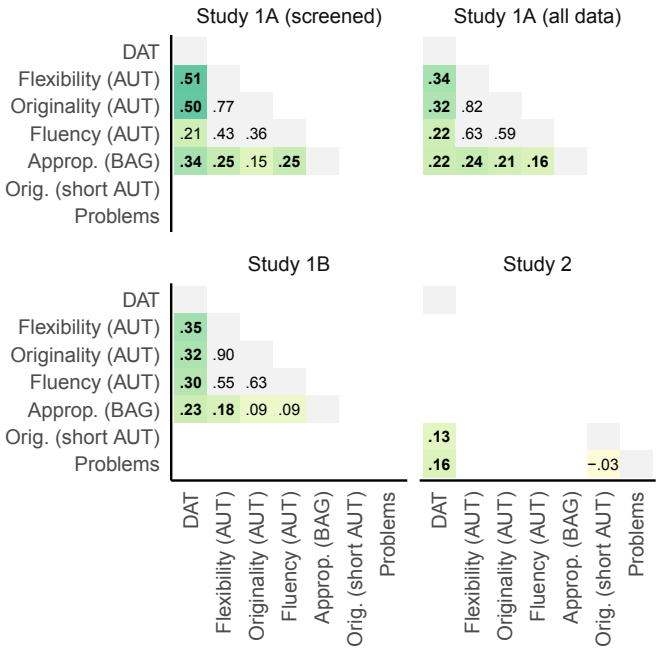

Fig. 4. Correlations across samples. The DAT showed the strongest correlations (colored columns) compared with the other measures (colored rows). Bold text shows between-task correlations greater than zero based on a one-tailed test. Within-task correlations are shown in gray. AUT, Alternative Uses Task; BAG, Bridge-the-Associative-Gap Task; Approp., Appropriateness; Orig., Originality.

feedback of the participant's score (https://osf.io/4kxh3/), allowing it to be used in contexts such as education or workshops. Enjoyment, brevity, and feedback are especially important for online tasks with reduced or no compensation, in which intrinsic motivation plays a central role in the desire to participate.

Limitations. The DAT has several limitations. Naming unrelated words measures originality with better face validity than appropriateness, although both are important components of creativity (31). DAT scores may thus partly reflect other constructs more related to divergence than creativity, such as overinclusive thinking or schizotypy (46). Still, the DAT scores correlated with assessments of appropriateness on the Bridge-the-Associative-Gap Task. Another limitation is that participants may artificially modulate their scores using different strategies to generate the words. Intentionally choosing rare words, looking around the room for inspiration, and following letter-based strategies (e.g., choosing rhyming words) can all influence the overall score. The fairly short time limit of 4 min may reduce the likelihood that participants will consider and implement these various strategies. Beyond the task itself, a limitation of our study is the small number of measures used. Given the high correlation between originality and flexibility in the Alternative Uses Task scores and the low number of items in the problem-solving tasks, future studies could replicate our results using other scoring methods as well as longer and more diverse creativity measures.

Future Research. Like all computational scoring methods, the DAT depends on the model and corpus used. We chose the GloVe algorithm and the Common Crawl corpus; this combination correlates best with human judgments on the Alternative Uses Task (27). For simplicity, we chose a pretrained model that is freely available (35). With some effort, researchers could train models using corpora from different countries at different times. As particular word associations become more or less related, the updated models would automatically account for these changes. This would allow the DAT scores to reflect the fluctuations in the cultural lexicon, as events and popular media shape word usage. Training models with corpora from different languages, as has been done with other algorithms (47), could even allow for global assessments of creativity. Still, for simplicity and ease of comparison, we recommend that researchers begin with the model tested here.

Future research could also use the DAT in experimental contexts requiring simple and short tasks. Since the DAT requires brief single-word responses, it may be suited for neuroimaging contexts in which movement must be minimized. Its simplicity may also make it suitable for completion in altered states of consciousness conducive to divergent thinking, such as presleep states or when using psychedelic drugs (48, 49). Research could also examine how different contexts, or changes to the wording of the instructions (50), influence word choices and the resulting DAT scores. Finally, studies could assess discriminant and criterion validity in more detail by examining a wider range of related and unrelated measures. Related measures could include memory retrieval ability, fluid intelligence, or real-world creative achievement (1, 2, 51, 52).

In sum, we demonstrate that naming unrelated words can be a reliable measure of divergent thinking. We hope this finding provides researchers with a simpler method of collecting and scoring creativity data across larger and more diverse samples.

Materials and Methods

Study 1. We validated the DAT against two well-established measures of creativity: the Alternative Uses Task and the Bridge-the-Associative-Gap Task. Study 1 contains three samples of participants who completed a demographics questionnaire and then several creativity tasks on a computer. All studies were approved by the University of Melbourne Human Research Ethics Committee (1954931.1), and all participants gave informed consent.

Participants. We recruited undergraduates from psychology courses in Melbourne, Australia (Studies 1A and 1B) and through social media advertisements in Montreal, Canada (Study 1C). In each sample, most participants were women and between 18 and 20 y old (Table 1).

Materials.

Alternative Uses Task. In the Alternative Uses Task, participants were presented with common objects in a random order: a brick, paper clip, newspaper, ice tray, and rubber band (15). The participants were instructed: "This is a test of creativity. Please list as many (and varied) uses as you can." Participants had 2 min to respond to each item in a text box; with five items, the task took up to 10 min.

Two independent raters later scored the responses using a uniqueness method adapted from DeYoung et al. (16). Originality scores depend on the frequency of the response within the sample. Participants received zero points for responses given by over 10% of the sample (e.g., using a rubber band to tie up hair), one point for responses given by 3 to 10% (e.g., as a string), two points for those given by 1 to 3% (e.g., as a tourniquet), and three points for responses given by under 1% (e.g., as dental floss). Flexibility scores, which often correlate with originality, reflect the number of different categories of uses mentioned. Using a rubber band as an eraser or to grip bottles would represent two distinct categories; using it to tie plastic bags or to group wires would represent one category (tying things together). Fluency is simply the number of distinct responses given in 2 min. We averaged each measure across the items for each participant. Across the samples, the two raters showed high interrater reliability (flexibility:

Table 1. Participant demographics across samples

	Study				
	1A (full)	1B	1C	2	Total
Country	Australia	Australia	Canada	98 countries	98 countries
N	141	285	50	8,572	8,914
Female, %	82	68	76	59	59
Age *M*, y	20.12	19.22	20.84	43.51	42.59
Age SD, y	4.05	2.17	2.68	17.66	17.93
Age range, y	18–47	16–47	18–33	6–70	6–70

In Study 2, ages are approximate since they were reported in bins.

$r = 0.94$ to 0.97; originality: $r = 0.64$ to 0.89; fluency: $r = 0.99$ to 1.00), so we averaged their scores.

Bridge-the-Associative-Gap Task. In the Bridge-the-Associative-Gap Task, a test of convergent thinking, participants were presented with pairs of words that were either related or unrelated to each other. Participants were asked to write a third word that is semantically related to both of the words. For example, if presented with giraffe and scarf, participants could write neck as the third word. The participants were given 30 s to respond to each item. In Study 1A, we randomly selected 20 of each type of pairs (related or unrelated) from the original set (36). Study 1B used the entire set.

Two judges then assessed the appropriateness of each response from one to five based on whether it related to both words in the pair. For example, a response of neck would be judged as appropriate (5) given giraffe and scarf, but a response of cheese would not (1). The judges generally agreed on their ratings ($r = 0.67$ to 0.78), so we averaged their scores.

DAT. Participants were asked to generate 10 unrelated nouns (*SI Appendix*, section S2 and https://osf.io/bxjhc/files have pencil-and-paper versions of the task). The task had the following additional instructions.

1) Use only single words.
   We used this rule because computational methods can score single words with less ambiguity than phrases. Words such as "cul de sac" were accepted and automatically hyphenated.
2) Use only nouns (e.g., things, objects, concepts).
   This rule keeps the class of words similar since the distance between words varies based on their part of speech, such as whether they are nouns or adjectives.
3) Avoid proper nouns (e.g., no specific people or places).
4) Avoid specialized vocabulary (e.g., no technical terms).
   This rule and the previous one prevent participants from using words that are too specific, which is one strategy to artificially inflate the score. To enforce these rules, only lowercase words from a common dictionary (53) were used in the calculation.
5) Think of the words on your own (e.g., do not just look at objects in your surroundings).
   During pilot testing, many participants would look around their environment for inspiration when naming the words. This strategy resulted in lower scores since common objects on one's desk are often semantically similar.
6) You will have 4 min to complete this task.
   In our initial testing (54), this amount of time was sufficient to complete the task without much time pressure.

After the task, participants were asked what strategy they used, if any, to choose the words. In Study 1A, two raters coded the responses based on 1) whether the 141 participants appeared to correctly follow the instructions and 2) whether they reported implementing a strategy such as naming the objects around them. Disagreements were resolved by discussion, and raters were liberal with their exclusions. Overall, 57 participants appeared to follow the instructions and not use a strategy; we used this manually screened subset for part of our analyses. In Studies 1B and 1C, we did not use this manual procedure to keep the scoring entirely automated.

Analysis. To test the relationship between the creativity measures, we checked for nonlinearity and then did one-tailed tests of linear correlation, with an α of 0.05 and no familywise type I error correction. All assumptions were reasonable for the tests. In Studies 1A and 1B, we aimed to run at least 90 participants per sample, which gave 80% statistical power to detect medium correlations of $r = 0.3$.

Study 2. We also tested how DAT scores varied by age, gender, country, and languages spoken. We recruited a much larger and more diverse sample as part of a broader study on experiences reported during creativity and insight tasks.

Participants. Participants were recruited through television, radio, and social media advertisements as part of a campaign by the Australian Broadcasting Corporation. In total, 8,572 participants completed the study from 98 countries. Most of the participants were from Australia ($n = 4{,}770$) or the United Kingdom ($n = 615$). Participants reported ages in bins ranging from under 7 y old ($n = 6$) to 70 or over ($n = 963$), with most falling in the 35 to 54 age range ($n = 2{,}834$), making the sample older than the participants in Study 1. Again, the majority (59%) of the sample was female (Table 1).

Materials. Since participants did not receive compensation in Study 2, we used shorter versions of several of the creativity measures to keep the study length feasible (15 min).

Alternative Uses Task. We used a shortened version of the task in which participants were asked to generate a single new and imaginative use for each of two common household objects. The objects were randomly selected from a brick, rubber band, shoe, paper clip, cup, and ice tray. Given that participants generated a single use for each object, flexibility and fluency could not be evaluated, so we focused on originality. Two raters judged originality from one to five in a subsample of 389 participants. To reduce within-task variation, this subsample included all participants who were randomly assigned the most common two objects (here, a cup and an ice tray). As in the previous samples, the judges generally agreed on their ratings ($r = 0.66$), so we averaged their scores.

DAT. We used the standard version of this task, as in Study 1. Participants saw a 4-min timer while completing the task, but there was no time limit. Still, as in Study 1, they completed the task in approximately a minute and a half (*Mdn* = 95.66 s, SD = 53.04).

Problems. Participants then completed a total of four problems. Two items were from the Compound Remote Associates Test, a convergent thinking task commonly used to assess insight and creativity (9). In each trial, participants saw three cue words and tried to find a fourth word that formed a compound word with the cues. For example, given the words sense, courtesy, and place, participants would suggest the word common. The 2 trials were randomly selected from a larger set of 15 (9). Participants additionally completed 1 insight and 1 analytical problem taken from a larger set of 20 [tables 4 and 5 in Webb et al. (37)]. An example insight problem is as follows: "How much earth is there in a hole 3 m by 3 m by 3 m?" It is often accompanied by a sudden feeling of clarity. An example analytical problem is as follows: "Using only 7-min and 11-min hourglasses, how can you time exactly 15 min?"

Two raters independently judged the accuracy of the answers on each of the four trials. A third rater resolved answers judged as ambiguous. Reliability between the two raters was high ($r = 0.87$); we considered answers as correct only when both raters scored them as such. Participants then reported their feelings during each problem (e.g., pleasure) from zero (nothing) to five (strong) as part of a larger study.

Demographic information. We collected brief demographic information aimed at maintaining anonymity: age (in bins), gender, country, and whether the participant was multilingual (i.e., whether they spoke a language other than English).

Procedure. On a website, participants were informed that the purpose of the study was to investigate their experiences with several creativity and problem-solving tasks. Participants then completed the shortened Alternative Uses Task, the DAT, and then the following items in a random order: two Compound Remote Associates Test items, one insight problem, and one analytical problem. For each measure, the items were randomly selected from a larger set to prevent participants from sharing answers with each other (e.g., if several people completed the task in the same room). All measures were in English, and there were no hard time limits. After the study, participants provided demographic information.

Analysis. All aspects of the study were preregistered online (https://osf.io/bfke8). To test the relationship between originality and the DAT scores, we checked for nonlinearity and then did a one-tailed test of linear correlation with an α of 0.05.

To assess how scores varied across basic demographics, we used ANOVA to test for main effects of 1) age (estimated by the minimum value of each age bin), 2) gender (female or male), 3) country (whether or not the participant was from Australia, the most common country in our sample), and 4) multilingualism (dichotomous), as well as interaction effects of 5) gender $\times$ age and 6) country $\times$ multilingualism. Using the Bonferroni correction, a familywise α of 0.10 gave a per-test α of 0.0167. Given our large expected sample size, we maintained high statistical power despite the low type I error rates.

Data Availability. The data and algorithm code have been deposited in the Open Science Framework (https://osf.io/vjazn/).

ACKNOWLEDGMENTS. We thank Elias Stengel-Eskin for help with the conceptualization of the task, as well as Victoria De Braga, Élissa Colucci, Marième Cyr, Ellen Langer, Daniel Little, and Claire Suisman for discussion and feedback. We are grateful for the assistance with data collection for Study 2 from Kylie Andrews, the Australian Broadcasting Corporation, and the National Science Week Committee. J.A.O. acknowledges funding from le Fonds de recherche du Québec–Santé. Study 1C was part of a larger project supported by Canada First Research Excellence Fund Grant 3c-KM-10 awarded to the Healthy Brains for Healthy Lives initiative at McGill University (Principal Investigator: Samuel Veissière).

PSYCHOLOGICAL AND COGNITIVE SCIENCES

1. M. Benedek *et al.*, How semantic memory structure and intelligence contribute to creative thought: A network science approach. *Think. Reas.* **23**, 158–183 (2017).
2. R. E. Beaty, P. J. Silvia, E. C. Nusbaum, E. Jauk, M. Benedek, The roles of associative and executive processes in creative cognition. *Mem. Cognit.* **42**, 1186–1197 (2014).
3. S. Mednick, The associative basis of the creative process. *Psychol. Rev.* **69**, 220–232 (1962).
4. Y. N. Kenett, D. Anaki, M. Faust, Investigating the structure of semantic networks in low and high creative persons. *Front. Hum. Neurosci.* **8**, 407 (2014).
5. E. Rossmann, A. Fink, Do creative people use shorter associative pathways? *Pers. Indiv. Differ.* **49**, 891–895 (2010).
6. Y. N. Kenett, M. Faust, A semantic network cartography of the creative mind. *Trends Cognit. Sci.* **23**, 271–274 (2019).
7. M. Benedek, A. C. Neubauer, Revisiting Mednick's model on creativity-related differences in associative hierarchies. Evidence for a common path to uncommon thought. *J. Creativ. Behav.* **47**, 273–289 (2013).
8. M. Becker, G. Wiedemann, S. Kühn, Quantifying insightful problem solving: A modified compound remote associates paradigm using lexical priming to parametrically modulate different sources of task difficulty. *Psychol. Res.* **84**, 528–545 (2018).
9. E. M. Bowden, M. Jung-Beeman, Normative data for 144 compound remote associate problems. *Behav. Res. Methods Instrum. Comput.* **35**, 634–639 (2003).
10. C.-L. Wu, H.-C. Chen, Normative data for Chinese compound remote associate problems. *Behav. Res. Methods* **49**, 2163–2172 (2017).
11. S. Acar, M. A. Runco, Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativ. Res. J.* **26**, 229–238 (2014).
12. M. Benedek, T. Könen, A. C. Neubauer, Associative abilities underlying creativity. *Psychol. Aesthet. Creat. Arts* **6**, 273–281 (2012).
13. R. W. Hass, Semantic search during divergent thinking. *Cognition* **166**, 344–357 (2017).
14. J. P. Guilford, Creativity. *Am. Psychol.* **5**, 444–454 (1950).
15. M. A. Wallach, N. Kogan, A new look at the creativity-intelligence distinction. *J. Pers.* **33**, 348–369 (1965).
16. C. G. DeYoung, J. L. Flanders, J. B. Peterson, Cognitive abilities involved in insight problem solving: An individual differences model. *Creativ. Res. J.* **20**, 278–290 (2008).
17. K. Beketayev, M. A. Runco, Scoring divergent thinking tests by computer with a semantics-based algorithm. *Eur. J. Psychol.* **12**, 210–220 (2016).
18. M. A. Runco, G. J. Jaeger, The standard definition of creativity. *Creativ. Res. J.* **24**, 92–96 (2012).
19. S. Acar, M. A. Runco, Divergent thinking: New methods, recent research, and extended theory. *Psychol. Aesthet. Creat. Arts* **13**, 153–158 (2019).
20. M. Benedek, C. Mühlmann, E. Jauk, A. C. Neubauer, 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. *Psychol. Aesthet. Creat. Arts* **7**, 341–349 (2013).
21. R. W. Hass, M. Rivera, P. J. Silvia, On the dependability and feasibility of layperson ratings of divergent thinking. *Front. Psychol.* **9** (2018), p. 1343.
22. J. A. Plucker, M. Qian, S. L. Schmalensee, Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking. *Creativ. Res. J.* **26**, 135–143 (2014).
23. R. Reiter-Palmon, B. Forthmann, B. Barbot, Scoring divergent thinking tests: A review and systematic framework. *Psychol. Aesthet. Creat. Arts* **13**, 144–152 (2019).
24. P. J. Silvia, Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Think. Skills Creativ.* **6**, 24–30 (2011).
25. P. J. Silvia *et al.*, Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychol. Aesthet. Creat. Arts* **2**, 68–85 (2008).
26. R. E. Beaty, D. R. Johnson, Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behav. Res. Methods* **53**, 757–780 (2021).
27. D. Dumas, P. Organisciak, M. Doherty, Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychol. Aesthet. Creat. Arts*, https://doi.org/10.1037/aca0000319 (2020).
28. R. Prabhakaran, A. E. Green, J. R. Gray, Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behav. Res. Methods* **46**, 641–659 (2013).
29. M. Benedek *et al.*, Creating metaphors: The neural basis of figurative language production. *Neuroimage* **90**, 99–106 (2014).
30. K. Gray *et al.*, "Forward flow": A new measure to quantify free thought and predict creativity. *Am. Psychol.* **74**, 539–554 (2019).
31. T. M. Amabile *et al.*, *Creativity in Context: Update to the Social Psychology of Creativity* (Routledge, 2018).
32. T. K. Landauer, D. Laham, B. Rehder, M. E. Schreiner, "How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans" in *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (Lawrence Erlbaum Associates, Mahwah, NJ 1997), pp. 412–417.
33. B. Forthmann, O. Oyebade, A. Ojo, F. Günther, H. Holling, Application of latent semantic analysis to divergent thinking is biased by elaboration. *J. Creativ. Behav.* **53**, 559–575 (2018).
34. F. J. Crosson, *Human and Artificial Intelligence* (Appleton-Century-Crofts, 1970).
35. J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation" in *Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA 2014), pp. 1532–1543.
36. L. R. R. Gianotti, C. Mohr, D. Pizzagalli, D. Lehmann, P. Brugger, Associative processing and paranormal belief. *Psychiatr. Clin. Neurosci.* **55**, 595–603 (2001).
37. M. E. Webb, D. R. Little, S. J. Cropper, Once more with feeling: Normative data for the aha experience in insight and noninsight problems. *Behav. Res. Methods* **50**, 2035–2056 (2018).
38. J. A. Olson, D. A. Sandra, D. Chmoulevitch, A. Raz, S. P. L. Veissière, A ten-step behavioural intervention to reduce screen time and problematic smartphone use. https://psyarxiv.com/tjynk/ (4 January 2021).
39. C. Stevenson *et al.*, Automated AUT scoring using a big data variant of the consensual assessment technique. https://modelingcreativity.org/blog/wp-content/uploads/2020/07/ABBAS_report_200711_final.pdf. Accessed 9 July 2020.
40. J. Baer, J. C. Kaufman, Gender differences in creativity. *J. Creativ. Behav.* **42**, 75–105 (2008).
41. H. W. Reese, L.-J. Lee, S. H. Cohen, J. M. Puckett, Effects of intellectual variables, age, and gender on divergent thinking in adulthood. *Int. J. Behav. Dev.* **25**, 491–500 (2001).
42. A. Abraham, Gender and creativity: An overview of psychological and neuroscientific literature. *Brain Imag. Behav.* **10**, 609–618 (2016).
43. A. Adnan, R. Beaty, P. Silvia, R. N. Spreng, G. R. Turner, Creative aging: Functional brain networks associated with divergent thinking in older and younger adults. *Neurobiol. Aging* **75**, 150–158 (2016).
44. M. Palmiero, D. D. Giacomo, D. Passafiume, Divergent thinking and age-related changes. *Creativ. Res. J.* **26**, 456–460 (2014).
45. M. A. Runco, S. Acar, Do tests of divergent thinking have an experiential bias? *Psychol. Aesthet. Creat. Arts* **4**, 144–148 (2010).
46. S. Kyaga *et al.*, Mental illness, suicide and creativity: 40-year prospective total population study. *J. Psychiatr. Res.* **47**, 83–90 (2013).
47. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages. https://arxiv.org/pdf/1802.06893.pdf (28 March 2018).
48. A. Mavromatis, *Hypnagogia: The Unique State of Consciousness between Wakefulness and Sleep* (Routledge and Kegan Paul, 1987).
49. L. Prochazkova *et al.*, Exploring the effect of microdosing psychedelics on creativity in an open-label natural setting. *Psychopharmacology* **235**, 3401–3413 (2018).
50. P. R. Christensen, J. P. Guilford, R. C. Wilson, Relations of creative responses to working time and instructions. *J. Exp. Psychol.* **53**, 82–88 (1957).
51. C. S. Lee, D. J. Therriault, The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. *Intelligence* **41**, 306–320 (2013).
52. S. H. Carson, J. B. Peterson, D. M. Higgins, Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativ. Res. J.* **17**, 37–50 (2005).
53. L. Németh, Hunspell. https://hunspell.github.io/. Accessed 26 September 2020.
54. J. A. Olson, L. Suissa-Rocheleau, M. Lifshitz, A. Raz, S. P. L. Veissière, Tripping on nothing: Placebo psychedelics and contextual factors. *Psychopharmacology* **237**, 1371–1382 (2020).