Semantic Distance and the Alternate Uses Task:

Recommendations for Reliable Automated Assessment of Originality

Roger E. Beaty[1], Dan R. Johnson[2], Daniel C. Zeitlen[1], & Boris Forthmann[3]

[1]Department of Psychology, Pennsylvania State University

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

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

Author Note

All files for analyses are available at Open Science Framework: https://osf.io/96zge/?view_only=33f2c20f962d4a7d95543f35c8b8baa3.

R.B. is supported by a grant from the National Science Foundation [DRL-1920653]. This research was supported by grant RFP-15-12 to R.B. from the Imagination Institute (www.imagination-institute.org), funded by the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the view of the Imagination Institute or the John Templeton Foundation.

Correspondence should be addressed to Roger Beaty, Department of Psychology, 140 Moore Building, University Park, PA, 16801, USA; rebeaty@psu.edu.

Abstract

Semantic distance is increasingly used for automated scoring of originality on divergent thinking tasks, such as the Alternate Uses Task (AUT). Despite some psychometric support for semantic distance—including positive correlations with human creativity ratings—additional work is needed to optimize its reliability and validity, including identifying maximally reliable items (objects) for AUT administration. We identify a set of 13 AUT items based on a systematic item-selection strategy (*belt, brick, broom, bucket, candle, clock, comb, knife, lamp, pencil, pillow, purse, sock*). This item-set resulted in acceptable reliability estimates and was found to be moderately related to both human creativity ratings and a creative personality factor (Study 1). These results replicated in a new sample of participants (Study 2). We conclude with the following recommendations for reliable and valid assessment of AUT originality using semantic distance: 1) make choices based on theoretical/practical considerations, 2) administer (some or all of) the 13 items from this study; 3) if other items must be used, avoid compound words as AUT items (e.g., guitar string); 4) include as many AUT items as time permits; 5) instruct participants to "be creative"; and 6) address fluency confounds that conflate idea quantity and quality (e.g., via max scoring).

Keywords: Alternate Uses Task; Short Form; Semantic Distance; Reliability; Validity

Semantic Distance and the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality

Creativity researchers are increasingly using computational tools such as semantic distance to assess creativity. Semantic distance provides an automated alternative to manual scoring by human raters, which is inherently labor-intensive and subjective. Perhaps the most common application of semantic distance has been to the alternate uses task (AUT)—a widely used measure of divergent thinking that involves producing creative uses for objects (Guilford, 1967). Despite the promise of semantic distance for automating AUT originality scoring, researchers have begun to identify some of its limitations, such as producing scores that are confounded by more elaborate responses (Forthmann et al., 2019; Forthmann, Holling, Çelik, et al., 2017) and the presence of "meaningless" stop words (Forthmann et al., 2019; Hass, 2017). In the present research, we aim to further improve upon the psychometrics of semantic distance in the context of the AUT by focusing on item characteristics (i.e., AUT objects; e.g., brick, rope), a largely unexplored but potentially critical feature of automated creativity assessment that likely impacts reliability and validity (Forthmann et al., 2016). We conduct two studies to identify a set of AUT items that the creativity community can use in future research on divergent thinking assessment with semantic distance.

Scoring Creativity with Semantic Distance

The question of how to best quantify the creative quality of ideas is a longstanding topic in creativity research (Amabile, 1982; Forthmann, 2019; Reiter-Palmon et al., 2019; Silvia et al., 2008; Wilson et al., 1953). One popular method has been to simply ask other people what they think: present a long list of ideas, e.g., uses for objects on the AUT, to minimally-trained human raters and ask them to rate the ideas on an ordered categorical scale (e.g., 1 = *not at all creative*, 5 = *very creative*; (Silvia et al., 2008). This approach, known as the subjective scoring method, has proven to be remarkably effective. There is now considerable evidence for the reliability and validity of subjective creativity scoring, including studies showing moderate to large correlations between human ratings on the AUT and real-world creative achievement

(Jauk et al., 2014). Despite its strengths, subjective scoring has its limits; most notably, human raters don't always agree on what they find creative, and they are often asked to score thousands of responses—leading to rater fatigue and negatively impacting the reliability of their ratings (Forthmann, Holling, Zandi, et al., 2017).

To address the limitations of subjective scoring and other manual methods, researchers are increasingly employing automated approaches, such as semantic distance. Semantic distance captures the originality (or novelty) facet of creative thinking by quantifying conceptual dissimilarity—the extent to which concepts are "far apart" from each other. The use of semantic distance in creativity research is based on the classic associative theory (Mednick, 1962), the notion that creative thinking involves making connections between remotely associated concepts. Early applications of semantic distance in creativity studies used a method called latent semantic analysis (LSA; (Landauer et al., 1998; Landauer & Dumais, 1997) to compute the semantic distance between words on verbal creativity tasks (e.g., the AUT; (Guilford, 1967; Wallach & Kogan, 1965). LSA is a form of distributional semantics that quantifies relationships between words in large corpora of natural language texts, such as books and other literary works, by computing the cosine similarity between word vectors in a high dimensional space. Words that tend to occur in similar contexts also have a higher similarity value (or lower distance values). For example, the words *pen* and *paper* tend to occur in the same contexts and would thus have a low semantic distance value; in contrast, the words *pen* and *boat* tend to occur in dissimilar contexts and would thus have a high semantic distance value.

Semantic distance has shown encouraging evidence of reliability and validity in studies on divergent thinking, word association, forward flow (i.e., the degree of change within a stream of thought), and the remote associates test (Beaty et al., 2014, 2021; Beisemann et al., 2019; Gray et al., 2019; Heinen & Johnson, 2018; Prabhakaran et al., 2014). For example, Prabhakaran, Green, and Gray (2014) applied semantic distance to the verb generation task, which presents nouns and asks participants to "think creatively" when generating verbs that can be associated with them. Participants who produced more semantically-distant verb associations

(as assessed by LSA) also performed better on established creativity tests (the Torrance Test and a creative writing test) and they reported more creative achievements. Regarding divergent thinking, recent studies have reported positive correlations between AUT semantic distance and measures of creative personality and achievement—openness to experience, creative self-efficacy, and creative activities/accomplishments (Beaty & Johnson, 2021; Dumas, Doherty, et al., 2020; Dumas, Organisciak, et al., 2020)—supporting the validity of semantic distance in divergent thinking assessment. Furthermore, Heinen and Johnson (2018) found that, when instructed to "think creatively" during verb generation, people spontaneously considered both novelty and appropriateness when generating responses: compared to instructions to think of "common" or "random" responses, instructions to think creatively yielded semantic distance scores between the extremes of common (least distant) and random (most distant). The authors also found that semantic distance scores correlated most strongly with human ratings of originality (compared to creativity and appropriateness).

Another increasingly popular application of semantic distance is forward flow (FF), a chained free association task which quantifies "how far people travel" in their stream of thought—or how much current thoughts diverge from preceding thoughts—via the semantic distance between word associations (i.e., semantic evolution; Gray et al., 2019). Gray and colleagues (2019) provided evidence of the reliability and validity of FF scores (as assessed by LSA), showing that FF scores robustly predict creativity. Specifically, the researchers found that FF was positively associated with several measures of creative thinking and creative behavior/achievement across different samples in laboratory and real-world settings, including a positive correlation with human AUT ratings even when controlling for general cognitive ability. These findings illustrate how the application of semantic distance to cognitive tasks beyond divergent thinking can also be useful in creativity research.

In a recent study (Beaty & Johnson, 2021), we sought to build upon the LSA findings by Prabhakaran et al. (2014) and others by expanding the computational models used to compute semantic distance, including 1) multiple machine learning models that use prediction

methods to estimate word similarity (including counting co-occurrences, e.g., LSA and GloVe) and 2) newer text corpora that leverage naturalistic language (such as subtitles from movies; Beaty & Johnson, 2021; cf. Dumas et al., 2020). Like human raters, semantic models have different "opinions" about novelty—due in part to variability in text corpora (e.g., textbooks vs. movie subtitles)—and estimating semantic distance from many different spaces should yield a composite value that is more generalizable than a single model alone. We conducted five studies to validate this multi-model approach to semantic distance computation for divergent thinking assessment. When applied to the AUT—where semantic distance was computed between the AUT object (e.g., sock) and participant responses (e.g., filtration device)—we found large latent correlations between semantic distance and human creativity ratings. Similar to Prabhakaran and others, we also found that people who produced more semantically-distant AUT responses also tended to report more creative activities and achievements (assessed via the Biographical Inventory of Creative Behaviors, Creative Achievement Questionnaire, and Inventory of Creative Activities and Accomplishments), as well as higher levels of openness to experience and more creative self-efficacy, providing additional evidence that semantic distance offers a valid and automated alternative to human creativity ratings.

Another recent study explored the reliability and validity of the multi-model approach developed by Beaty and Johnson (2021) in the context of forward flow assessment. By averaging FF scores across seven semantic spaces, Beaty et al. (2021) showed that a multi-model approach yielded increased reliability of FF scores compared to LSA only. Furthermore, Beaty and colleagues created a latent FF factor with the averaged FF scores as indicators, and found that FF also predicted human creativity ratings of AUT responses, even when controlling for intelligence in a structural equation model. Altogether, these studies provide encouraging evidence on the psychometric properties of multi-model approaches to measuring semantic distance in both divergent thinking and free association tasks.

In addition to the psychometric strengths of semantic distance, researchers are beginning to identify some limitations. For example, Forthmann et al. (Forthmann et al., 2019;

Forthmann, Holling, Çelik, et al., 2017) found that “additive” LSA compositional models (which add word vectors when computing semantic distance scores) can be confounded by more elaborate AUT responses. That is, simply having more words in an AUT response systematically influenced the LSA distance values. However, Forthmann and colleagues found that this elaboration bias could be attenuated by removing “stop words" (or “meaningless” words; e.g., he, have, me, the, them) from AUT responses. Other recent work has sought to test different compositional models (e.g., multiplying word vectors) to determine which yields the most reliable and valid semantic distance values with AUT responses (Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Maio et al., 2020). In our view, such psychometric studies are critical to further strengthen the semantic distance approach for reliable and valid creativity assessment.

The Present Research

Semantic distance is a promising automated method for scoring verbal creativity tasks, with increasing evidence of its reliability and validity (Beaty et al., 2021; Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Maio et al., 2020). Yet recent work has also identified aspects of divergent thinking tasks that significantly impact the scores produced by semantic distance algorithms (e.g., elaboration bias; Forthmann et al., 2019; Maio et al., 2020). In the present research, we explore another potential influential source of variability in semantic distance scores on the AUT. Specifically, we examine whether different AUT objects (e.g., box, rope, brick) yield different semantic distance values, and whether some objects perform better than others. Currently, researchers commonly include one or two AUT objects, computing the semantic distance of responses as the outcome measure. However, this approach assumes that all items equally measure the construct of interest (divergent thinking), despite several researchers employing multi-object paradigms to reduce item-specificity in studies using conventional scoring metrics (e.g., fluency and originality; Barbot, 2018; Kleinkorres et al., 2021; Wilken et al., 2019). Indeed, AUT items have been shown to have varying item

characteristics that can be partially explained by psycholinguistic variables, such as object frequency (Forthmann et al., 2016).

We therefore conducted two studies to identify a list of items that measure divergent thinking on the AUT reliably and validly with semantic distance. Borrowing from the neuroscience literature on divergent thinking, which typically presents many short trials to isolate the (neural) signal of interest (Benedek et al., 2019), as well as recent work emphasizing the merits of including multiple trials in idea generation tasks (instead of one or two; Barbot, 2018; Kleinkorres et al., 2021), we sought to construct a version of the AUT that consists of several short trials. To this end, we leveraged an existing dataset of AUT items and participant responses (Study 1), and conducted item analyses and factor analyses to determine which items load onto a coherent latent semantic distance factor. We then examined the construct validity of this approach by assessing how the resulting semantic distance values correlate with human creativity ratings (Study 1) and creative personality (i.e., openness to experience, creative self-efficacy, and creative behavior; Studies 1 and 2). In sum, we sought to bolster the psychometric integrity of semantic distance for verbal creativity assessment, producing a "short form" of the AUT that can reliably and validly measure divergent thinking.

Study 1

Our first study reanalyzed data from a recent fMRI study that included several AUT items (Beaty et al., 2018). For each item, participants generated a single creative use, which was subsequently analyzed via semantic distance. We employed a factor analytic approach to determine which of the 46 items loaded onto a common latent factor. Given past work reporting stimulus effects for divergent thinking studies using fluency and originality indices (Barbot, 2018; Forthmann et al., 2016; Kleinkorres et al., 2021), along with our recent observations of variable inter-item correlations using semantic distance (Beaty et al., 2021), we hypothesized that the factor analysis would yield a limited number of AUT items that load highly and significantly on the semantic distance factor. To assess the validity of this approach, we computed correlations between the semantic distance factor and 1) human creativity ratings of

the same AUT items and 2) creative personality and behavior measures (openness, creative self-efficacy, and creative activities).

Method

Participants

Data for Study 1 were collected as part of a larger project on the neuroscience of creativity and imagination (see Beaty et al., 2018). The full sample of participants consisted of 186 adults from the University of North Carolina at Greensboro (UNCG) and the surrounding community; participants who completed the MRI phase of the larger project were included in the present analysis (n = 175; 129 women, mean age = 22.74 years, SD = 6.37). Participants completed written consent forms and received up to $100 for their participation in the multiphase study, which consisted of neuroimaging (see Beaty et al., 2018), cognitive assessment (for more details, see (Frith et al., 2021) and daily-life experience sampling (see Zeitlen et al., 2021). Consistent with common inclusion criteria for MRI research, participants were right-handed, with normal or corrected-to-normal vision, and they reported no history of cognitive impairment, neurological issues, or drugs affecting the central nervous system. The study procedure was approved by the UNCG Institutional Review Board (IRB).

Procedure

Participants first completed the MRI session, which lasted approximately one hour. They then completed a one-hour battery of cognitive assessments and personality scales on a desktop computer running MediaLab experiment software.

Divergent Thinking Assessment

Participants completed two tasks during fMRI in an event-related design: 1) the AUT and 2) the Object Characteristics Task (OCT; see Beaty et al., 2018). They were presented with a series of 46 objects (see Appendix A). Most of the 46 objects were derived from previous fMRI studies (e.g., (Fink et al., 2009), and others were generated by the authors based on face validity/perceived conduciveness to generating uses. Items were randomly assigned to the AUT and OCT conditions within-person. Importantly, this approach yielded highly sparse coverage

for the 46 items across the sample (i.e., all participants completed 23 AUT trials, but the AUT objects they used varied). For the AUT, participants were asked to think of a single creative use for each object; if they had an idea before the thinking period expired, they were encouraged to continue thinking of the most creative idea they could. Participants were explicitly instructed to "think creatively" (Acar et al., 2020) and to try to come up with the most original idea they could during the thinking period. The OCT is a common semantic control task in fMRI studies of divergent thinking (Beaty et al., 2015; Fink et al., 2009), and it requires participants to think of the defining physical features of a series of objects (23 trials); OCT responses were not analyzed in the current study (Beaty et al., 2018). The fMRI trial structure consisted of: (a) a jittered fixation cross (4-6 s), (b), a condition cue (3 s), (c) a silent response generation phase (12 s), and (d) a response production phase, during which participants spoke their response into an MRI-compatible microphone (5 s; cf., Beaty et al., 2017; Benedek et al., 2014). Before the fMRI scanning session, participants received thorough instructions and completed several practice trials. Verbal responses were transcribed by an experimenter for subsequent assessment of creative quality by four trained raters using the subjective scoring method (Silvia et al., 2008). Raters rated each response on a 5-point scale, from 1 (*not at all creative*) to 5 (*very creative*); they were trained to assess responses on three dimensions: uncommonness, remoteness, and cleverness (cf., Wilson et al., 1953).

Semantic Distance Computation

AUT responses were also coded for semantic distance using the *SemDis* platform (semdis.wlu.psu.edu; Beaty & Johnson, 2021). For each response, semantic distance was computed for five semantic models: two count models and three predict models. Count models (e.g., LSA) count the co-occurrences of words in text corpora; predict models (e.g., word2vec) try to predict a given word from surrounding context words using machine learning. The two count models were: 1) a latent semantic analysis (LSA) model, Touchstone Applied Science Associates (TASA), which computes word co-occurrences within a text corpus (37,651 documents, middle and high school textbooks and literary words, 92,393 different words),

followed by a singular value decomposition on the resulting sparse matrix (300 dimensions; Günther et al., 2015; cf., Prabhakaran et al., 2014); and 2) the global vectors (GloVe; Pennington et al., 2014) model, which is trained on ~6 billion tokens (300 dimensions, top 400,000 words) and uses weighted least squares to extract global information across a concatenation of the 2014 Wikipedia dump and the Gigaword corpus (news publications from 2009-2010). The three predict models were: 1) a concatenation of the ukwac web crawling corpus (~ 2 billion words) and the subtitle corpus (~385 million words; window size = 12 words, 300 dimensions, most frequent 150,000 words; Mandera et al., 2017); 2) the subtitle corpus only (window size 12 words, 300 dimensions, most frequent 150,000 words); and 3) a concatenation of the British National Corpus (~2 billion words), ukwac corpus, and the 2009 Wikipedia dump (~ 800 million tokens; window size = 11 words, 400 dimensions, most frequent 300,000 words; Baroni et al., 2014).

Prior to computing semantic distance, responses were preprocessed using the "remove filler and clean" setting on the *SemDis* platform. This automated approach removes "stop words" (e.g., the, an, a, to) and punctuation marks that can confound semantic distance computation (Forthmann et al., 2019; Forthmann, Holling, Çelik, et al., 2017). For all five semantic models, we computed the semantic distance between the AUT object (e.g., pencil) and participants' responses using the "all" semantic space setting on *SemDis*. Finally, we selected the "multiplicative" compositional model option on *SemDis* to account for AUT responses with multiple words. Multiplicative models multiply word vectors that are computed for each response (i.e., semantic distance between the AUT object and all words in a given response; (Beaty & Johnson, 2021). For all analyses, we used the average semantic distance of the five models; if words were not found in a given semantic space (which occurred rarely in both samples), the value was missing, and the average was computed from the available models (Beaty et al., 2021; Beaty & Johnson, 2021).

Personality Assessment

To validate the semantic distance approach, we included self-report scales to measure individual traits previously associated with creativity and AUT semantic distance, including openness to experience, creative self-concept, and creative behavior (Beaty & Johnson, 2021). Openness to experience was assessed using the Openness subscale of the Big Five Aspect Scale (BFAS), which includes 10 items such as "I need a creative outlet" (DeYoung et al., 2007). Creative self-concept was assessed using the Short Scale of Creative Self (Karwowski, 2011), which includes two facets (creative self-efficacy [6 items] and creative personality identify [5 items]), with items such as "I trust my creative abilities" (creative self-efficacy) and "I think I am a creative person" (creative personality identity). Creative behavior was assessed using the Biographical Inventory of Creative Behaviors (Batey, 2007; Silvia et al., 2021), which asks people to indicate whether they have engaged in 34 creative behaviors in the last 12 months (yes/no response), such as writing a short story, organizing an event, and making a present.

Data Analysis

We used the statistical software R (R Core Team, 2019) to perform all reported analyses in this work. All files for analyses are available in the Open Science Framework (OSF; https://osf.io/96zge/?view_only=33f2c20f962d4a7d95543f35c8b8baa3). We used the R package mice (van Buuren & Groothuis-Oudshoorn, 2011) to handle missing data by means of multiple imputation. In addition, we used the packages miceadds (Robitzsch & Grund, 2021), psych (Revelle, 2020), lavaan (Rosseel, 2012), and semTools (Jorgensen et al., 2021).

Several challenges were inherent in the analysis of this dataset: a) there was substantial missing data, b) we sought to select the best candidate AUT items, and c) we aimed to perform an initial reliability and validity evaluation of the item set. The missing data pattern was Missing Completely at Random (MCAR) because of the random assignment of objects to persons. Hence, using multiple imputation was well justified. We used a total of $m = 40$ imputed datasets following suggestions in the literature (Azur et al., 2011). However, given that the data matrix was quite sparse (range of percentages of missing values across AUT objects was from 46.29% to 65.14%) which potentially leads to underestimated covariances (cf. Hardt et al.,

2012), we complemented our strategy by analyzing correlation matrices based on pairwise deletion. Pairwise deletion is unbiased when missing values are MCAR (Newman, 2014), but it is well-known that pairwise deletion can yield non-positive definite correlation matrices. This issue was addressed by the cor.smooth() function of the psych package. Hence, we relied on both approaches to not limit our item-selection to only one potentially problematic strategy. We expected that this combined approach would result in a more robust set of items.

The goal was to develop a unidimensional semantic distance scale. Hence, we used Cureton's item-scale correlation that corrects for both part-whole overlap and unreliability of the scale composite (Cureton, 1966). We used item-scale correlation > .30 as an initial criterion for item selection. This criterion was applied to item-scale correlations derived from the pairwise-deletion approach. For the multiple imputation approach, the criterion was found empirically by checking the relationship between item-scale correlations based on both approaches (see Figure 1; for more details see below). After this initial item reduction step, the imputation approach was rerun based on the reduced item set to further stabilize the imputation approach (i.e., data matrix subjected to imputation was less sparse this way). A unidimensional CFA model was fit, and items were further scrutinized for potentially displaying residual covariances with other items (based on modification indices > 5; see Jöreskog & Sörbom, 1988). Such items were further excluded to prevent a bulky correlational structure.

We further employed multiple imputation for validity examination. For each validity criterion, a separate imputation was run to reduce complexity of the analysis. First, 40 datasets were imputed based on SemDis scores and human ratings for the selected AUT objects to correlate SemDis scores with human ratings. We used passive imputation (van Buuren & Groothuis-Oudshoorn, 2011) to impute average scores across all items for both SemDis scores and human ratings. This approach was chosen to further reduce the complexity of this validity check (i.e., we considered latent variable modeling here as being too complex and opted for the more pragmatic approach). The correlation was pooled across imputed datasets by means of the micombine.cor() function from the miceadds package. This function also provides a confidence

interval and *p* value. The validity evaluation with creative personality as a criterion was based either on a multiple imputation approach (40 imputed datasets, including all thirteen AUT objects and the five creative personality indicators) or pairwise deletion correlations. Latent variable modeling was then used based either on the multiple imputation or the pairwise deletion covariance matrix.

Results

Item Selection

As expected, inter-item covariances were on average smaller for the imputation approach as compared to the pairwise deletion approach. However, we found a correlation near unity between item-scale correlations based on either imputation or pairwise deletion (see Figure 1). Hence, we decided to use a criterion > .30 for pairwise deletion item-scale correlations. This corresponded with a criterion of > .21 for imputation item-scale correlations (inferred based on linear regression of imputation item-scale correlations on pairwise deletion item-scale correlations; see Figure 1). Consequently, we selected items with item-scale correlations > .30 for pairwise deletion and > .21 for imputation. Applying these criteria resulted in an initial set of 20 items (see Table A1 in the Appendix A). It is further noteworthy that two items only passed the criterion for multiple imputation, but not for pairwise deletion. This observation highlights the usefulness of complementing both missing data handling strategies.

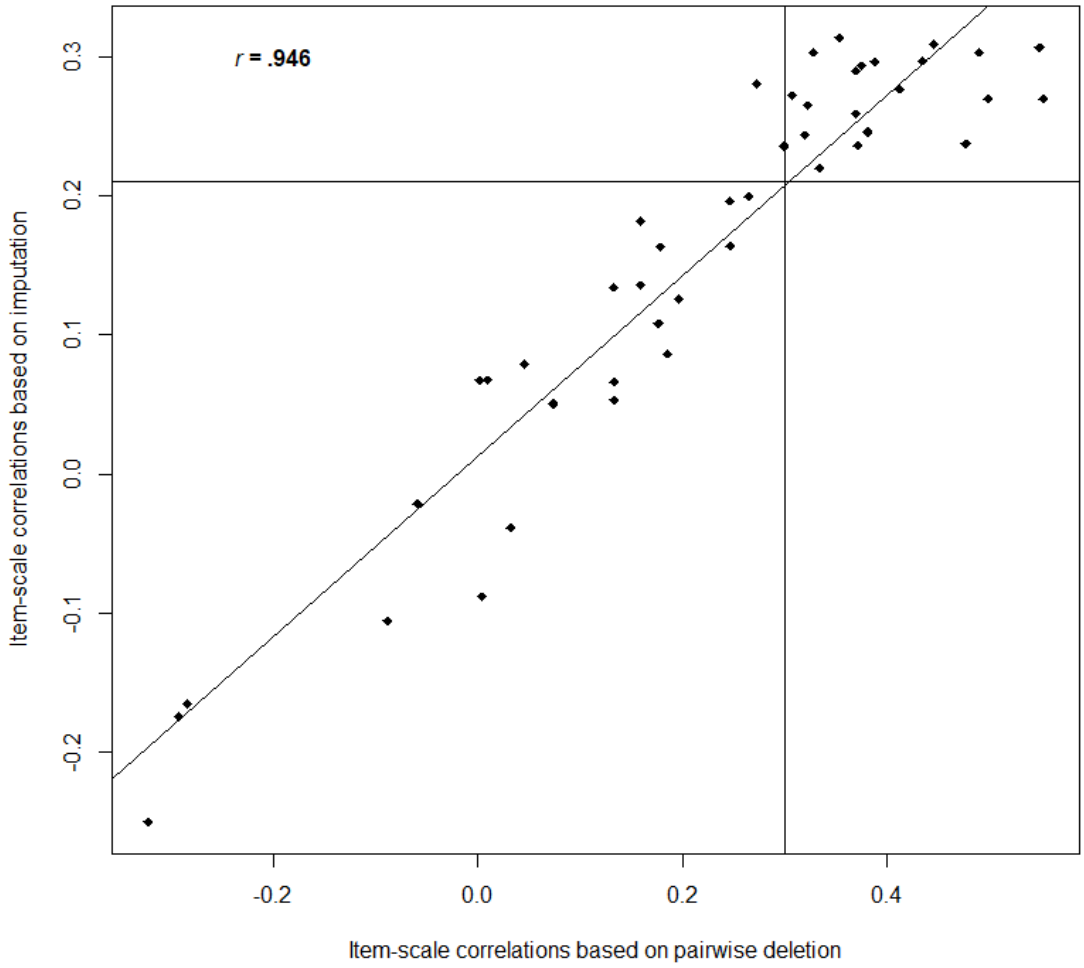
In a next item selection step, modification indices for residual covariances between items were inspected for estimated unidimensional models based on both missing data handling approaches. We further excluded seven items that displayed significant residual covariances based on modification indices > 5 (see Table A1 in the Appendix A and Table B1 in Appendix B).

The selected items for the preliminary SemDis scale consisted of the items *belt*, *brick*, *broom*, *bucket*, *candle*, *clock*, *comb*, *knife*, *lamp*, *pencil*, *pillow*, *purse*, and *sock* (see Figure 2 and Table A1 in the Appendix A1). The model fit for a unidimensional CFA model was acceptable only for SRMR and $\hat{\gamma}$ based on the MI approach (see Table 1) and for the χ^2/df ratio

based on the PD approach. The low values obtained for CFI and TLI were explainable based on the small values for the RMSEA of the null model (regardless of the missing data handling approach; see Table 1).

Figure 1

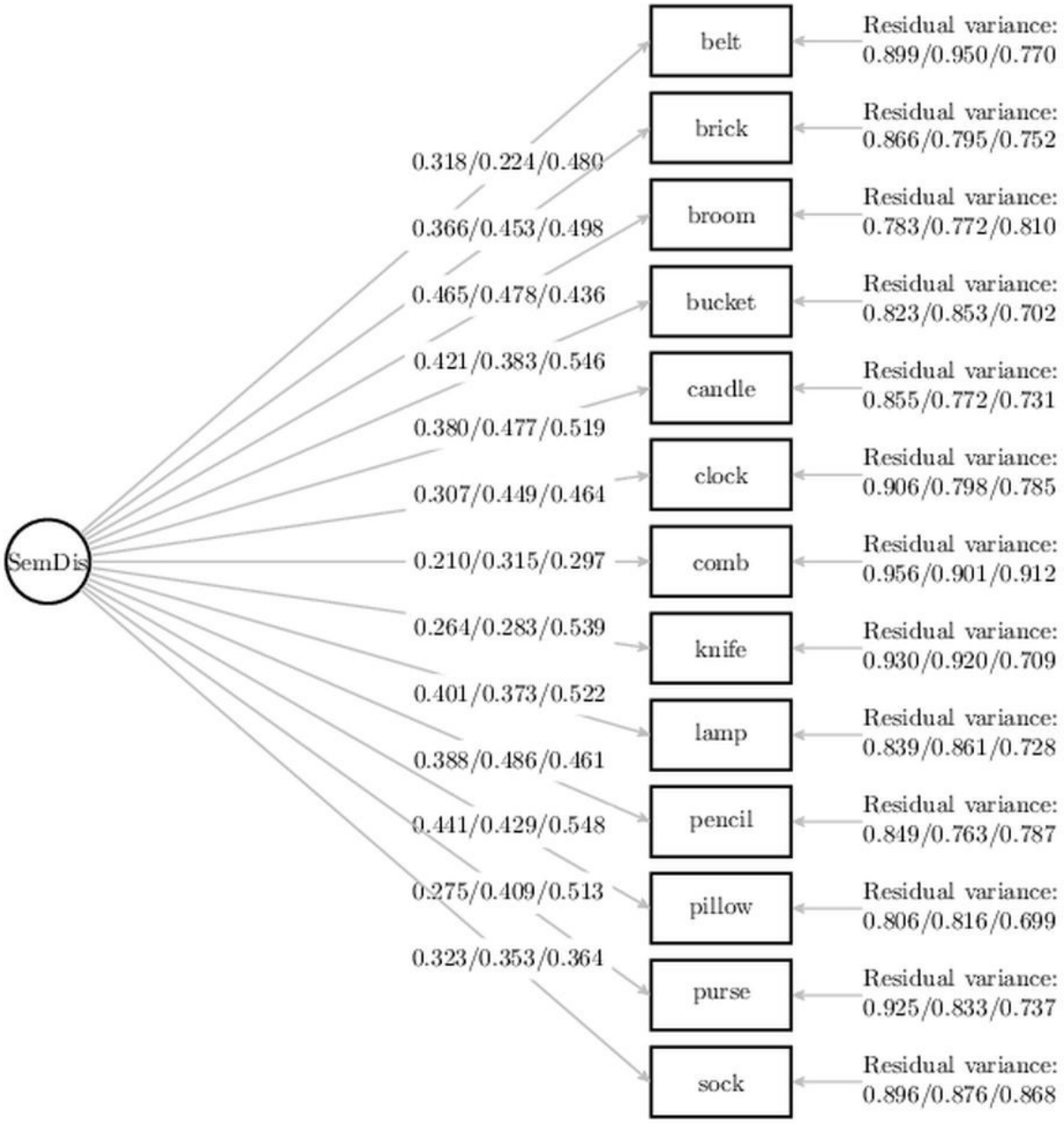
Item-scale correlations based on imputation and pairwise deletion



Notes. The regression line of imputation-based item-scale correlations on pairwise deletion item-scale correlations is depicted. The vertical and horizontal lines represent the cut-offs used for item selection.

Figure 2

Unidimensional CFA results from Study 1 and Study 2



Notes. Depicted are standardized estimates. Results are separated by a forward slash and reported in the following order: Study 1 – imputation-based/Study 1 – based on pairwise deletion/Study 2. Manifest variables are represented by rectangles. Latent variables are represented by circles.

Table 1

CFA model fit for unidimensional SemDis models and two-dimensional models for SemDis and creative personality

	Unidimensional SemDis model[a]			Two-dimensional validity model[b]		
Fit index	Study 1 – MI	Study 1 – PD	Study 2	Study 1 – MI	Study 1 – PD	Study 2
$\chi^2(df)$	145.77 (65)***	123.13 (65)***	77.76 (65)	337.36 (103)***	389.57 (103)***	112.07 (103)
χ^2/df	2.24	1.89	1.20	3.28	3.78	1.09
RMSEA	.084	.108	.036	.111	.180	.024
SRMR	.079	.106	.063	.093	.125	.061
CFI	.583	.535	.939	.515	.322	.968
TLI	.499	.443	.927	.435	.210	.962
$\hat{\gamma}$	.934	.896	.987	.864	.706	.992
RMSEA – null model	.119	.144	.148	.147	.202	.134

Notes. MI = results are based on a multiple imputation covariance matrix. PD = results are based on a smoothed pairwise deletion correlation matrix. [a]The unidimensional SemDis model is depicted in Figure 2. [b]The two-dimensional validity includes one SemDis latent variable and one creative personality latent variable.

Preliminary Reliability Findings

SemDis scoring resulted in roughly acceptable levels of reliability (see Table 2). The estimates based on the multiple imputation approach were somewhat lower as compared to the pairwise deletion approach.

Table 2

Reliability results

	Cronbach's α	Coefficient ω_1
Study 1 – MI	.638	.641
Study 1 – PD	.702	.706
Study 2	.785	.790

Notes. MI = results are based on a multiple imputation covariance matrix. PD = results are based on a smoothed pairwise deletion correlation matrix. Coefficient ω_1 is a reliability estimate based on structural equation modeling that does not assume essential τ-equivalence (Bollen, 1980; Raykov, 2001).

Preliminary Validity Findings

The average SemDis score across the selected items correlated moderately with average human ratings, $r = .351$, 95%-CI: [.125, .543], $p = .003$.

Next, we assessed the correlation between SemDis and creative personality latent variables. The measurement model for creative personality included three observed variables and, hence, the unidimensional model was a saturated model that cannot be evaluated by classical fit indices. However, substantial standardized loadings (range from .47 to .75) were found across indicators and both approaches (i.e., MI and PD). The fit of the two-dimensional model, including a SemDis and a creative personality latent variable, was close to acceptable levels only for very few indicators; fit was clearly better based on the MI-approach (see Table 1). The latent variable correlation between SemDis and creative personality was significantly positive and small to moderate in size (this correlation varied only negligibly across MI and PD approaches to handle missing data).

Table 3

Validity Results

Criterion	SemDis	*r*	*p*	95%-CI
Study 1				
Human ratings	Average Score[a]	.351	.003	[.125, .543]
Creative Personality	Latent variable – MI	.291	.004	[.095, .486]
Creative Personality	Latent variable – PD	.299	.033	[.023, .574]
Study 2				
Creative Personality	Latent variable – FIML	.195	.077	[-.021, .411]
Integrated analysis				
Creative Personality	Latent variable – weighted average[b]	.248	.001	[.103, .393]

Notes. MI = latent variable model was based on multiple imputation covariance matrix. PD = latent variable model was based on a smoothed pairwise-deletion correlation matrix. [a]Average scores for human ratings and SemDis were based on a passive imputation approach. [b]The correlation between SemDis and creative personality were averaged by means of the R package metafor (Viechtbauer, 2010). We used the MI-based correlation from Study 1; the integrated analysis differed only slightly when performed with the PD-based correlation.

Discussion

Study 1 resulted in a reasonable set of 13 candidate AUT items with promising psychometric properties. Reliability estimates were roughly acceptable across the used approaches to handle missing data. In particular, for the MI approach, it can be reasonably assumed that the reliability estimates of around .65 represented underestimates. Given that reliability estimates used in this work rely on covariance-based methods, in combination with the fact that such covariances are susceptible to underestimation in MI involving many variables (i.e., rather scarce data matrices), it is reasonable to assume that reliability is underestimated for this approach. Reliability estimates based on the PD approach reached acceptable levels and,

hence, we conclude that SemDis scoring based on the 13-item candidate set is promising in terms of reliability.

Notably, the RMSEA of the null model was smaller than .158 for both approaches, and it is well known that CFI and TLI cannot be higher than their recommended cut-offs in this case (see the webpage of David A. Kenny for a discussion of this issue: http://davidakenny.net/cm/fit.htm). It should be further mentioned that $\hat{\gamma}$ is not affected by this issue. Taking this into account, we conclude that the unidimensional SemDis model based on the MI-approach yielded acceptable model fit. However, the same conclusion cannot be drawn for all models (i.e., the unidimensional SemDis model and the two-dimensional model involving SemDis and creative personality) estimated based on the PD-approach. Despite these identified technical issues, we found that factor loadings (see Figure 2) and also validity results (see Table 3) based on latent variable modeling yielded highly comparable findings.

In addition, the correlation between SemDis and human ratings of creative quality was moderate in size. To reduce complexity of the analysis, this correlation was based on average observed scores derived from a passive imputation approach. Hence, this correlation is not corrected for measurement error and, thus, validity in this regard is expected to be stronger as compared to the found estimate of .351.

Interestingly, none of the compound items (e.g., guitar string) were selected by the employed item selection strategy. Some of the compounds even exhibited negative item-total correlations. This could perhaps be explained by technical problems introduced by the fact that compounds are represented by a two-word vector. While previous work on the SemDis approach (Beaty & Johnson, 2021) has shown that multiplicative composition of word vectors (compared to additive composition) can successfully suppress problems arising from an elaboration bias (Forthmann et al., 2019), this might still show up when items vary in terms of word length. We recommend assessing this issue more closely in future studies and recommend refraining from using compound AUT items when employing the SemDis approach.

Study 1 leveraged a large item sample of AUT objects. Hence, the preliminary item set for a SemDis scale is based on a very broad initial item sample. On the contrary, several technical issues were identified as a consequence of the quite sparse data matrix resulting from random item selection (a feature of the fMRI task design; Beaty et al., 2018). We employed a cautious strategy based on two different approaches to handle missing data and our item selection strategy highlighted that the strategies complemented each other well. However, given that some issues with latent variable modeling and potential underestimation of inter-item correlations remained, we sought to replicate these findings in a second study.

Study 2

Study 1 identified a set of AUT object cues that loaded onto a latent semantic distance factor. Importantly, 13 out of the 46 items showed significant loadings, indicating that not all objects are treated equally with semantic distance computation. It is worth noting that Study 1 used data from an fMRI study, where the AUT was administered in an atypical context that required participants to generate a single idea (with 12 seconds to think and 5 seconds to speak). This design has the virtue of controlling fluency, which can confound divergent thinking assessment (Forthmann, Szardenings, & Dumas, 2020; Forthmann, Szardenings, & Holling, 2020).

In Study 2, we aimed to replicate and extend the findings of Study 1 by administering the 13 items in a typical testing environment (i.e., on desktop computers). In addition, Study 1 had some notable technical issues, and given that it was not clear if the unidimensional model replicates well in a second study, we additionally assessed four more items for the case of unsuccessful replication. However, to accommodate all 17 items in a reasonable amount of time, we gave participants 30 seconds to generate responses. Although conventional testing time ranges from 2-3 minutes, Study 1 indicates that reliable and valid individual differences can be distilled from very brief idea generation windows (e.g., 12 seconds). To account for variation in fluency, as was done in Study 1, we used the max scoring approach, taking the highest semantic distance value per each AUT item. Max scoring has been shown to be most promising in terms

of validity of semantic distance (Forthmann et al., 2019). To validate this approach, we again aspects of creative personality (i.e., openness, creative self-efficacy, and creative behavior).

Method

Participants

Study 2 data were collected as part of a study on verbal creativity assessment. The full sample of participants consisted of 151 adults from Penn State University (PSU; 100 women, mean age = 19.31 years, *SD* = 1.79). Participants completed consent forms and received credit toward a research option in a psychology course. The study was approved by the PSU IRB.

Procedure

Participants completed a battery of cognitive assessments and personality scales using the online experiment platform Pavlovia. They were asked to complete the online study in a quiet room with minimal distractions.

Divergent Thinking Assessment

Participants completed the AUT using 17 items from Study 1. They were given 30 seconds to think of (and type) creative object uses; the "thinking time" in this study was greater than Study 1, which was constrained by the short trial durations required in fMRI studies (Benedek et al. 2019). The instructions were similar to those used in Study 1, and they were consistent with our past work on divergent thinking that emphasize creativity (Silvia et al., 2008): participants were asked to "think creatively" and "to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different." The order of AUT trials was randomized for each participant.

Responses were scored via SemDis using the same approach as Study 1. To account for fluency confounds—variability in the number of responses for each participant, which biases summed originality values (Forthmann, Szardenings, & Holling, 2020)—the max scoring approach was employed. Specifically, for each of the 17 AUT trials, the most semantically-distant response was selected and included in subsequent reliability and validity analyses.

Personality Assessment

To validate semantic distance scores, the same "creative personality" scales from Study 1 were administered (with the exception of a different openness scale): NEO FFI Openness (12 items; McCrae & Costa, Jr., 2007), Short Scale of Creative Self (creative self-efficacy and creative personal identity), and Biographical Inventory of Creative Behaviors.

Data Analysis

Data analysis was again performed by means of the statistical software R (R Core Team) and its lavaan package (Rosseel, 2012). The range of missing values across the thirteen items was from 1.34% to 9.40% and the assumption of Missing Completely at Random could not be refuted based on Jamshidian et al.'s (2014) two-step procedure (Hawkins test of normality and homoscedasticity: $p < .001$; non-parametric test of homoscedasticity: $p = .141$). In addition, the three creative personality indicators had only 1.32% missing values, and Jamshidian et al.'s MCAR test revealed again that the MCAR assumption could not be refuted (Hawkins test of normality and homoscedasticity: $p < .001$; non-parametric test of homoscedasticity: $p = .126$). Hence, we used full information maximum likelihood to handle missing data. Given that multivariate normality was clearly violated (only SemDis scores: $b_{1,13} = 76.15$; $A = 1294.54$; $p < .001$, $b_{2,13} = 264.40$; $B = 17.23$; $p < .001$; SemDis and creative personality scores: $b_{1,16} = 99.29$; $A = 1687.94$; $p < .001$, $b_{2,16} = 349.37$; $B = 12.91$; $p < .001$), we also used robust Maximum Likelihood estimation. Fit indices were the same as in Study 1.

Results

The unidimensional CFA model displayed excellent model fit in Study 2 (see Table 1). Again, the null model RMSEA was quite small (see Table 1) which implies that CFI and TLI results should be interpreted with caution. The standardized factor loadings ranged from .297 to .548 (see Figure 2) and reliability estimates were clearly acceptable and higher as compared to Study 1 (see Table 2).

In addition, the two-dimensional model, including a SemDis and a creative personality latent variable, displayed excellent fit to the data (see Table 1). The correlation between both variables was found to be small and significant by trend (see Table 3). Finally, to integrate

validity findings across Study 1 and Study 2, we used the rma() function of the metafor package (Viechtbauer, 2010) and weighted the average correlation by the respective squared standard errors. This highlights an overall small to moderate significant correlation between SemDis and creative personality.

Discussion

Study 2 aimed at replicating and extending the findings of Study 1. Importantly, the unidimensionality of the SemDis scale replicated, and the model fit and reliability estimates were stronger as compared to Study 1. The positive correlation between SemDis and creative personality, however, was found to be somewhat smaller than Study 1. An integrated analysis—aggregating results from Study 1 and 2—revealed a moderately positive correlation between the SemDis and creative personality factors.

General Discussion

Creativity researchers are increasingly using automated approaches to assess originality on divergent thinking tasks, such as semantic distance, addressing the subjectivity and labor cost of subjective/manual scoring methods. The present research sought to improve upon the psychometric properties of semantic distance, which has shown sensitivity to particular features of divergent thinking tasks and responses (Forthmann et al., 2019). We conducted two studies to examine a potentially important but under-studied feature of the AUT: item characteristics (i.e., the objects that people use to think of creative uses).

Study 1 leveraged a large set of AUT items from a recent fMRI study on divergent thinking (Beaty et al., 2018), finding a reduced set of 13 items that yielded acceptable reliability (assessed by semantic distance) and correlated positively with human creativity ratings and a creative personality factor, providing validity evidence. Study 2 replicated the reliability findings of Study 1, showing an attenuated correlation between SemDis scores and creative personality that, taken together with Study 1 via an integrated analysis (i.e., combining results from both studies), yielded a moderately positive correlation. Our findings suggest that not all

AUT items are treated equally by semantic distance algorithms, but that reliable and valid semantic distance scores can be obtained by using the 13 items identified in this work.

Our study provides a set of AUT items that can be used in behavioral research using semantic distance. However, neuroimaging experiments, particularly fMRI, requires many more items/trials for reliable neural measurement (Benedek et al., 2019). This issue could be addressed by pooling AUT stimuli across studies/labs[1], and conducting further psychometric analysis to identify a larger set of reliable items suitable for fMRI research (tested under brief/single response generation conditions typical to the fMRI environment). We recommend not including compound words (e.g., guitar string) if responses will be analyzed using semantic distance, given the problematic values they yielded in Study 1.

As it remains unclear to what extent the effects of item characteristics observed in the present study are specific to the AUT or divergent thinking tasks, we encourage future studies to explore how item characteristics may influence the psychometric properties of other cognitive tasks scored automatically using semantic distance (e.g., forward flow). Another important task feature in divergent thinking assessment is the time allowed for idea generation (Paek et al., 2021) because idea quality usually increases with more time on task (Acar et al., 2019; Bai, Leseman, et al., 2021; Bai, Mulder, et al., 2021; Hass, 2017). Our studies used very brief idea generation periods (12s in Study 1, 30s in Study 2), due to time constraints of fMRI (Study 1) and to limit participant fatigue when administering several AUT items (Study 2). There is indeed evidence for consistency of creative performance across varying time conditions (Forthmann, Lips, Szardenings, et al., 2020), but whether these observations extrapolate to the time limits used in this work is yet to be determined. Hence, future work is needed to identify optimal time limits for semantic distanced-based originality scoring on the AUT, as has been done with manual/subjective originality scoring (Benedek et al., 2013; Paek et al., 2021). Additionally, future studies should further assess the impacts of item word length and elaboration bias. However, human raters also tend to rate longer responses as more creative

[1] We thank an anonymous reviewer for this suggestion.

(Beaty & Johnson, 2021), so elaboration bias may be a concern for both automated and subjective scoring of originality in divergent thinking responses.

Based on the present findings, as well as related psychometric work on semantic distance, we propose the following recommendations to promote reliable and valid assessment of originality using semantic distance and the AUT:

1. *Make decisions based on theoretical/practical considerations.* As it is the case for divergent thinking assessment in general (Reiter-Palmon et al., 2019), theoretical deliberations and/or the purpose of assessment should guide any choices with respect to semantic distance scoring of divergent thinking tasks. The recommendations presented here are applicable to a wide range of research purposes, but they may require adaption for very specific research questions, e.g., altering task instructions (see point 5) to study individuals with a high need for uniqueness, to assess whether these individuals show different originality scores when presented with different types of task instructions.
2. *Administer (some or all of) the 13 items identified in this work* (we tested 46 items in Study 1, and only these 13 were found to be reliable with a simple unidimensional structure). If using a subset, check psychometric properties using the openly available data for Study 2 (e.g., reliability of 4 items).
3. *If other AUT items must be used, avoid compound items* (e.g., guitar string), which showed highly problematic psychometric features (i.e., item-scale correlations).
4. *Include as many AUT items as time permits* to limit item-specific effects (cf. Barbot, 2018; Kleinkorres et al., 2021; Wilken et al., 2019).
5. *Instruct participants to "be creative"* (cf. Acar et al., 2020; Said-Metwaly et al., 2020). This should be the default instruction for divergent thinking assessment because it is most transparent for participants. If participants are not told to think creatively when completing divergent thinking tasks, and responses are then scored for creativity/originality, a mismatch in task instructions and scoring threatens

reliability and validity. In some cases, other instructions (e.g., "be fluent") may be considered if they are of theoretical interest. Recommended "be creative" instructional language can be found here: https://osf.io/vky36/.

6. *Address fluency confounds that conflate idea quantity and quality*, e.g., via max scoring (Forthmann et al., 2019), top scoring (Hass, 2017), average scoring (Beaty & Johnson, 2021; Dumas, Organisciak, et al., 2020; Dumas & Dunbar, 2014), or fixing the number of responses that participants are asked to produce for each trial (e.g., 2-3; Barbot, 2018; Zarnegar et al., 1988). Decisions to use a specific scoring method depend on practical/theoretical considerations (see point 1). For example, if the equal odds baseline is tested, average scoring should be used and, depending on the statistical approach, even a sum score can be useful. No other originality scoring fits the statistical framework of the equal odds baseline.

References

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

Acar, S., Abdulla Alabbasi, A. M., Runco, M. A., & Beketayev, K. (2019). Latency as a predictor of originality in divergent thinking. *Thinking Skills and Creativity*, *33*, 100574. https://doi.org/10.1016/j.tsc.2019.100574

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

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bai, H., Leseman, P. P. M., Moerbeek, M., Kroesbergen, E. H., & Mulder, H. (2021). Serial Order Effect in Divergent Thinking in Five- to Six-Year-Olds: Individual Differences as Related to Executive Functions. *Journal of Intelligence*, *9*(2), 20. https://doi.org/10.3390/jintelligence9020020

Bai, H., Mulder, H., Moerbeek, M., Kroesbergen, E. H., & Leseman, P. P. M. (2021). Divergent thinking in four-year-old children: An analysis of thinking processes in performing the Alternative Uses Task. *Thinking Skills and Creativity*, *40*, 100814. https://doi.org/10.1016/j.tsc.2021.100814

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

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Batey, M. (2007). *A psychometric investigation of everyday creativity*. University of Londong.

Beaty, R. E., Benedek, M., Barry Kaufman, S., & Silvia, P. J. (2015). Default and Executive Network Coupling Supports Creative Idea Production. *Scientific Reports*, *5*(1), 10964. https://doi.org/10.1038/srep10964

Beaty, R. E., Christensen, A. P., Benedek, M., Silvia, P. J., & Schacter, D. L. (2017). Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *NeuroImage*, *148*, 189–196. https://doi.org/10.1016/j.neuroimage.2017.01.012

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., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapil, T. R., Kane, M. J., & Silvia, P. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, *115*(5), 1087–1092. https://doi.org/10.1073/pnas.1713532115

Beaty, R. E., Silvia, P., Nusbaum, E., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197. https://doi.org/https://doi.org/10.3758/s13421-014-0428-8

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

Beisemann, M., Forthmann, B., Bürkner, P.-C., & Holling, H. (2019). Psychometric Evaluation of an Alternate Scoring for the Remote Associates Test. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.394

Benedek, M., Christensen, A. P., Fink, A., & Beaty, R. E. (2019). Creativity assessment in neuroscience research. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 218–226. https://doi.org/10.1037/aca0000215

Benedek, M., Jauk, E., Fink, A., Koschutnig, K., Reishofer, G., Ebner, F., & Neubauer, A. C. (2014). To create or to recall? Neural mechanisms underlying the generation of creative new ideas. *NeuroImage*, *88*, 125–133. https://doi.org/10.1016/j.neuroimage.2013.11.021

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

Bollen, K. A. (1980). Issues in the Comparative Measurement of Political Democracy. *American Sociological Review*, *45*(3), 370. https://doi.org/10.2307/2095172

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

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., & 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

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000319

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., & Neubauer, A. C. (2009). The creative brain: Investigation of brain activity during creative problem solving by means of EEG and FMRI. *Human Brain Mapping*, *30*(3), 734–748. https://doi.org/10.1002/hbm.20538

Forthmann, B. (2019). Die Beurteilung von Ideenqualität. In J. S. Haager & T. G. Baudson (Eds.), *Kreativität in der Schule - finden, fördern, leben* (pp. 75–95). Springer. https://doi.org/10.1007/978-3-658-22970-2_4

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

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

Forthmann, B., Lips, C., Szardenings, C., Scharfen, J., & Holling, H. (2020). Are Speedy Brains Needed when Divergent Thinking is Speeded—or Unspeeded? *The Journal of Creative Behavior*, *54*(1), 123–133. https://doi.org/10.1002/jocb.350

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of Latent Semantic Analysis to Divergent Thinking is Biased by Elaboration. *The Journal of Creative Behavior*, *53*(4), 559–575. https://doi.org/10.1002/jocb.240

Forthmann, B., Szardenings, C., & Dumas, D. (2020). On the Conceptual Overlap between the Fluency Contamination Effect in Divergent Thinking Scores and the Chance View on Scientific Creativity. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.445

Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of

fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 94–112. https://doi.org/10.1037/aca0000196

Frith, E., Kane, M. J., Welhaf, M. S., Christensen, A. P., Silvia, P. J., & Beaty, R. E. (2021). Keeping Creativity under Control: Contributions of Attention Control and Fluid Intelligence to Divergent Thinking. *Creativity Research Journal*, *33*(2), 138–157. https://doi.org/10.1080/10400419.2020.1855906

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

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

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology*, *12*(1), 184. https://doi.org/10.1186/1471-2288-12-184

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

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 144–156. https://doi.org/10.1037/aca0000125

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech : An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR).

Journal of Statistical Software, *56*(6). https://doi.org/10.18637/jss.v056.i06

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

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL VII: A guide to the program and applications* (2nd ed.). SPSS.

Karwowski, M. (2011). It doesn't hurt to ask…But sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(2), 154–164. https://doi.org/10.1037/a0021427

Kleinkorres, R., Forthmann, B., & Holling, H. (2021). An experimental approach to investigate the involvement of cognitive load in divergent thinking. *Journal of Intelligence*, *9*(1). https://doi.org/10.3390/jintelligence9010003

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

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

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

McCrae, R. R., & Costa, Jr., P. T. (2007). Brief Versions of the NEO-PI-3. *Journal of Individual Differences*, *28*(3), 116–128. https://doi.org/10.1027/1614-0001.28.3.116

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3),

220–232. https://doi.org/10.1037/h0048850

Newman, D. A. (2014). Missing Data. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Paek, S. H., Abdulla, A., Acar, S., & Runco, M. A. (2021). Is More Time Better for Divergent Thinking? A Meta-Analysis of the Time-On-Task Effect on Divergent Thinking. *Thinking Skills and Creativity*, 100894. https://doi.org/10.1016/j.tsc.2021.100894

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, *46*(3), 641–659. https://doi.org/10.3758/s13428-013-0401-7

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 315–323. https://doi.org/10.1348/000711001159582

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

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

Silvia, P. J., Rodriguez, R. M., Beaty, R. E., Frith, E., Kaufman, J. C., Loprinzi, P., & Reiter-Palmon, R. (2021). Measuring everyday creativity: A Rasch model analysis of the Biographical Inventory of Creative Behaviors (BICB) scale. *Thinking Skills and Creativity*, *39*, 100797. https://doi.org/10.1016/j.tsc.2021.100797

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children; a study of the creativity-intelligence distinction*. Holt, Rinehart and Winston.

Wilken, A., Forthmann, B., & Holling, H. (2019). Instructions Moderate the Relationship between Creative Performance in Figural Divergent Thinking and Reasoning Capacity. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.392

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*(5), 362–370. https://doi.org/10.1037/h0060857

Zarnegar, Z., Hocevar, D., & Michael, W. B. (1988). Components of original thinking in gifted children. *Educational and Psychological Measurement*, *48*(1), 5–16. https://doi.org/10.1177/001316448804800103

Appendix A

This appendix includes Table A1 with all items shown that were initially studied.

Table A1

Alternate Uses Objects used for scale construction and item selection decisions

Object	First step	Second step	Included in SemDis scale
ashtray	Excluded	-	No
balloon	Selected	Excluded	No
baseball	Selected	Excluded	No
belt	**Selected**	**Selected**	**Yes**
book bag	Excluded	-	No
brick	**Selected**	**Selected**	**Yes**
broom	**Selected**	**Selected**	**Yes**
bucket	**Selected**	**Selected**	**Yes**
candle	**Selected**	**Selected**	**Yes**
CD	Excluded	-	No
clock	**Selected**	**Selected**	**Yes**
comb	**Selected**	**Selected**	**Yes**
dish	Excluded	-	No
dog leash	Excluded	-	No
drinking straw	Excluded	-	No
earring	Excluded	-	No
flower pot	Excluded	-	No
fork	Selected	Excluded	No
garden hose	Excluded	-	No
gas can	Excluded	-	No
guitar string	Excluded	-	No

gum wrapper	Excluded	-	No
hair dryer	Excluded	-	No
hanger	Excluded	-	No
hat	Selected	Excluded	No
kite	Excluded	-	No
knife	**Selected**	**Selected**	**Yes**
lamp	**Selected**	**Selected**	**Yes**
lighter	Selected	Excluded	No
newspaper	Excluded	-	No
pencil	**Selected**	**Selected**	**Yes**
pillow	**Selected**	**Selected**	**Yes**
plastic bag	Excluded	-	No
purse	**Selected**	**Selected**	**Yes**
razor	Selected	Excluded	No
ruler	Excluded	-	No
scarf	Selected	Excluded	No
screwdriver	Excluded	-	No
shoe	Excluded	-	No
shoelace	Excluded	-	No
shovel	Excluded	-	No
shower curtain	Excluded	-	No
soap	Excluded	-	No
sock	**Selected**	**Selected**	**Yes**
suitcase	Excluded	-	No

Notes. Selected items for the final SemDis scale are depicted in bold font. 'First step' item selection was based on Cureton's item-scale correlation calculated for both a multiple imputation covariance matrix and a smoothed pairwise deletion correlation matrix (see

Figure 1). ‘Second step’ item selection was based on modification indices > 5 for residual covariance parameters observed when a unidimensional CFA model was estimated based on either a multiple imputation covariance matrix and a smoothed pairwise deletion correlation matrix.

Appendix B

In this appendix we report all modification indices for residual covariances with values > 5 (Jöreskog & Sörbom, 1988) along with the respective object pairs (see Table B1).

Table B1

Residual Covariances with Modification Indices > 5

Object pair	Modification index	Standardized expected parameter change
MI approach		
Balloon – baseball	7.694	-.221
Baseball – broom	**5.525**	**.186**
Baseball – clock	**17.410**	**.329**
Baseball – hat	5.505	.183
Belt – hat	6.522	.204
Belt – knife	10.306	-.256
Belt – scarf	**15.530**	**.318**
Broom – scarf	**15.737**	**-.321**
Bucket – pillow	**8.007**	**.227**
Bucket – razor	5.845	-.198
Candle – lighter	**18.983**	**-.351**
Candle – pencil	5.355	-.187
Candle – scarf	**10.487**	**.261**
Candle – sock	**8.210**	**.227**
Clock – pencil	6.148	-.200
Comb – lighter	**7.823**	**.222**
Comb – pencil	7.060	.211
Knife – pillow	5.219	-.183
Knife – razor	**13.924**	**.303**

Knife – scarf	5.522	-.187
Lamp – pencil	5.251	.185
Lighter – razor	**8.340**	**.238**
Pillow – razor	**10.415**	**-.268**
PD approach		
Balloon – knife	8.188	.345
Baseball – broom	**5.088**	**.271**
Baseball – clock	**12.712**	**.433**
Belt – brick	5.267	.271
Belt – scarf	**11.320**	**.398**
Broom – scarf	**9.236**	**-.367**
Bucket – pillow	**6.056**	**.294**
Candle – lighter	**15.506**	**-.497**
Candle – scarf	**7.776**	**.335**
Candle – sock	**5.650**	**.282**
Comb – lighter	**6.244**	**.317**
Fork – hat	6.919	-.307
Fork – pencil	9.547	-.369
Hat – knife	7.436	-.321
Knife – pencil	5.536	.283
Knife – razor	**6.101**	**.299**
Lamp – razor	7.235	.328
Lighter – razor	**8.391**	**.373**
Pillow – razor	**9.105**	**-.371**

Notes. Variable pairs identified by both MI and PD approach are depicted in bold font.