

ORIGINAL ARTICLE

Psycholinguistics of divergent thinking prompts predict the originality of elementary students' responses

Theadora Rogue Vlaamster¹ , Selcuk Acar², Peter Organisciak³, Denis Dumas⁴ and Kelly Berthiaume²

¹Educational Psychology, Ball State Teachers College, USA, ²Educational Psychology, University of North Texas College of Arts and Sciences, USA, ³Education, University of Denver, USA and ⁴Educational Psychology, University of Georgia College of Arts and Sciences, USA

Corresponding author: Theadora Rogue Vlaamster; Email: theavlaamster@gmail.com

(Received 3 May 2025; revised 26 November 2025; accepted 4 March 2026)

Abstract

In the continued effort to identify creative talent, divergent thinking (DT) assessments are commonly employed to objectively assess aspects of individual creative ability. Despite their common usage, the selection of DT prompts remains unstandardized. This is concerning because it is unclear if all prompts can be employed equally or if certain characteristics associated with the acquisition and usage of said prompt in a given language affect the magnitude of response originality. Researchers administered a computer-based DT assessment to 386 elementary students and compared the originality scores of the respondents to the originality scores predicted by prompts' psycholinguistic features. Hierarchical linear modeling was employed to control individual DT ability and examine the effects of psycholinguistics on originality at two levels, with individual response originality scores (level 1) nested within individual respondents (level 2). Results indicated statistically significant main effects for word frequency, semantic diversity, prompt type, and item reaction time. There were also several significant interaction effects between prompt type and word frequency, prompt type and word length, and prompt type and semantic diversity. The resulting final model explained approximately 18% of the variance in originality scores, suggesting the need for further consideration of DT prompt choice.

Keywords: Creativity assessment; divergent thinking; hierarchical linear modeling; original thinking; psycholinguistics

Creativity, or the capacity to generate novel and useful ideas, solutions, or products (Runco & Jaeger, 2012; Stein, 1953), is a fundamental psychoeducational concept. Much effort has been dedicated to its assessment to identify and foster creative

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

talent (Kaufman et al., 2008; Runco & Acar, 2024b). Divergent thinking (DT) tasks have been the center of these efforts because they resemble key aspects of challenges that require creative thinking: they are ill-defined, open-ended, and devoid of a single correct solution. Whereas creativity is clearly broader than DT (Cropley, 2006), these tasks have been popular measures of creativity due to their objectivity compared to self-reports and greater association with measures of creative achievement than intelligence tests on average (Gerwig, et al., 2021; Kim, 2008).

DT tasks are either verbal or nonverbal (Plucker et al., 2024; Runco & Acar, 2024a). Examples of verbal DT tests include Alternate Uses (Guilford et al., 1958; “List different uses for a *brick*”), Instances (Wallach & Kogan, 1965; List things that *move on wheels*), Similarities (Wallach & Kogan, 1965; “How are carrots and refrigerator alike”), and Consequences (Guilford et al., 1958; What would happen if people had a sixth finger). These tasks present one or several verbal, open-ended prompt(s), instructing participants to generate many, different, and/or creative responses (see Acar et al. (2020) for various instructions used). Then, these responses are often scored for fluency (number of responses), flexibility (variety of responses), originality (unusualness, cleverness, and remoteness of responses), and elaboration (degree of detail and elegance of the responses), as well as sometimes for quality and appropriateness (usefulness).

Frequently, researchers select DT items without considering the nature of the prompts they selected. Because DT tasks are open-ended, the nature of the responses, which are used to obtain creative potential estimates, is potentially dependent on the prompts. In other words, the nature of the responses may be different when “cup” is selected as the Alternate Uses test prompt rather than “brick.” If this is true, such an item selection should be made deliberately, and factors influencing the scoring outcomes should be considered. Although this problem applies to any kind of DT test item, verbal DT tests allow objectively examining prompt-related factors due to the availability of psycholinguistic metrics of test prompts such as word frequency and age of acquisition (Forthmann et al., 2016; Ogurlu et al., 2023). This information is important because the selection of the tasks can be more deliberate. Some psychometric outcomes can be estimated before collecting the data, such as the reliability of the scores and the extent to which items are at the same difficulty level. This knowledge can also be used to determine the tasks used in pre- and post-assessments to ensure that at least some test prompts’ qualities are constant (Kaya & Acar, 2019). In this study, we focus on elementary school children’s responses to a DT test and examine if psycholinguistic characteristics of the verbal DT prompts can predict the originality of the responses.

Literature review

Extant literature has documented evidence linking psycholinguistic characteristics to aspects of creativity, such as ideational fluency. In one early study by Cofer and Shevitz (1952), who used word association tests (similar to DT tests in that both are open-ended), the number of associations made was positively related to the presented prompts’ word frequency count within the given language. In other words, prompts with greater frequency are linked to greater number of responses produced. Adapting

the Remote Associate Tests, Vitrano et al. (2021) tested the same effect by considering the stimuli's word frequency and concreteness (vs abstractness). They found a positive relationship between word frequency and the number of associations made, but there was no difference in terms of concreteness. Notably, these two tasks allow associations and thus provide a fluency score, but they are not true DT tasks.

Past research with fluency scores from DT tests provided somewhat mixed evidence. Forthmann et al. (2016) found that high-frequency objects generated more ideas than low-frequency ones. Additionally, they investigated the effect of explicit instructions (to be fluent versus to be creative) on fluency, revealing that high-frequency words led to significantly more ideas in the be-fluent condition than in the be-creative condition. However, no difference in fluency was observed for low-frequency words, regardless of the instructions. However, Forster (2008) reported no statistically significant relation between word frequency metrics and fluency in the Alternate Uses Test. However, words with more common uses were found to have lower originality. The inconsistent findings in Forthmann et al. (2016) and Forster (2008) on originality may be explained by different factors, such as how the DT tests were administered (explicit instructions, time; Acar et al., 2020; Paek et al., 2021), how word frequency was operationalized (lemma frequencies vs word form frequencies; Brysbaert & New, 2009), and linguistic and cultural differences (German vs. English).

Later, Ogurlu et al. (2023) conducted a meta-analysis of studies using Alternate Uses Test items and found that fluency is positively related to word frequencies when Alternate Uses Test items are administered with shorter time limits. In other words, using Alternate Uses Test items with higher word frequencies allows for more responses when these items are given with strict time limits. Ogurlu et al. (2023) explored some other psycholinguistic characteristics besides word frequency, such as word length, bigram, semantic diversity, lexical decision latency, and age of acquisition, and none of them were significant. Importantly, these psycholinguistic characteristics may predict performance on other aspects of DT, such as originality or flexibility, and such predictions can be made for different kinds of DT tests, such as Instances rather than Alternate Uses Test alone.

In the present work, we focus on originality and examine if the originality of produced responses is related to the psycholinguistic characteristics of their prompts. Wilson et al. (1953) defined original ideas as those that are unusual, clever, and remote. Such ideation can be effectively captured by DT tests, allowing for response generation with varying levels of originality. The importance of originality has been underlined in both empirical and conceptual work. Among different aspects of DT (e.g., fluency, flexibility, originality, and elaboration), originality (which is sometimes interchangeably used with novelty) is the one that is most strongly tied to creativity in empirical work (Acar et al., 2017; Diedrich et al., 2015), and it is also considered the backbone of creativity (Rothenberg & Hausman, 1976; Runco, 1988). Runco et al. (2005) stated that "originality is a necessary part of creativity, but creative things are more than just original" (p. 137).

Conceptual framework

Psycholinguistic characteristics of verbal DT prompts can predict the originality of the responses for various reasons. One key concept to link originality to

psycholinguistic characteristics is *prompt openness*. Ogurlu et al. (2023) defined openness as “the potential of an open-ended test prompt to lead to a greater or lower ideational fluency, originality, and flexibility due to the item characteristics” (p. 3). Runco and Albert (1985) argued that the level of openness of a DT item can enhance ideational originality due to a greater level of divergence and associative pathways. Openness can be seen in the figural DT test items as the degree to which the prompts resemble something specific or not (Runco et al., 2016). Figural prompts may be considered more open when they do not resemble anything (e.g., an object) specific. Witczak et al. (2024) compared performance on verbal Alternate Uses Test prompts to the performance when an image is added to the verbal prompt. This disambiguation led to higher fluency and lower originality, which is consistent with Runco and Albert’s (1985) prediction regarding originality. Verbal DT prompts may be considered more open depending on the degree to which the verbal prompt involves words or concepts that are easier to retrieve from memory and one’s familiarity with the prompt. Psycholinguistic characteristics can reflect and capture these characteristics of the prompts. Using words with a greater number of word frequency or semantic flexibility may then allow for more associative pathways to follow. According to this view, word frequency is a positive predictor of originality, like ideational fluency, as the verbal prompts provide more freedom for ideational exploration. This is also consistent with the extended effort hypothesis (Parnes, 1961) that views originality as a direct outcome of fluency. Past work has shown that word frequency is a positive predictor of originality, and the factors that enhance fluency could also enhance originality. In other words, selecting a highly open verbal DT item will likely trigger more responses (higher fluency), and having more ideas supply more original responses.

Another relevant concept is the path of least resistance (Finke et al., 1992). This theory suggests that individuals generate ideas that come to mind easily and often stop when generating ideas becomes more challenging. However, the most straightforward and most accessible ideas are usually the least creative, which also explains why responses generated for DT tasks are progressively more creative (Christensen et al., 1957). One could expect a negative relationship between word frequency and originality because prompts with a higher word frequency value may indicate lower “resistance” (higher accessibility), resulting in the generation of quickly accessible and ordinary responses. Supporting this point, Forster (2008) found that when prompt objects were familiar and lacked distinctive features, the originality of the responses decreased. Other factors, such as explicit instructions and respondents’ task commitment, can also play a major role. The potential negative impact of the path of least resistance on originality may be mitigated by explicit instructions or specific strategies such as brainstorming (Rietzschel et al., 2014), especially if the respondents aim to show their maximal performance.

Memory functions and experiences are at the nexus of the relationship between psycholinguistics and originality. On the one hand, memory is integral to forming psycholinguistic processes such as comprehension, production, and language acquisition (Harley, 2013; Menn & Dronkers, 2016). Many psycholinguistic metrics, such as the age of acquisition and lexical decision latency, rely on memory functions or its formation and development. On the other hand, overall, memory functions are positively related to creativity (Gerver et al., 2023; Madore et al., 2015).

Research showed that performance on the Alternate Uses Test is influenced by retrieval from episodic and semantic memory (Hass & Beaty, 2018) and broad retrieval ability (Miroshnik et al., 2023; Silvia et al., 2013). Accordingly, Runco and Acar (2010) found that original ideation benefits from both personal and social (vicarious) experiences. Conversely, Forster (2008) found a negative relationship between familiar prompt objects without distinctive features and the originality of the responses. Mixed findings regarding the impact of psycholinguistic characteristics on originality require further investigation. For example, Runco and Acar (2010) counted the number of original responses rather than assessing the average originality of the responses, which controls for fluency confound (Clark & Mirels, 1970). Thus, one key consideration is an assessment of originality that is not confounded by fluency. Likewise, Forster's (2008) findings relied on subjectively rated familiarity with the objects rather than a standard word frequency metric, although these two were correlated.

Psycholinguistic indicators

Psycholinguistic characteristics are broad, and some are more relevant to our investigation. One key characteristic is *word frequency*, which is defined as a word's frequency of occurrence in language. This is valuable because the frequency with which a word appears is one of the most reliable indicators of its processing efficiency (Brysbaert et al., 2018). Frequently occurring words within a language are more familiar; therefore, they are processed more quickly and recalled faster, while they may not be as easy to recognize as low-frequency words (Monsell et al., 1989; Yonelinas, 2002). Because DT items are open-ended and involve memory processes, word frequency metrics (e.g., British National Corpus Consortium, 2007; Brysbaert & New, 2009; Davies, 2008) may influence the quantity and quality of the responses (Dumas et al., 2021; Forthmann et al., 2016; Ogurlu et al., 2023).

Another key psycholinguistic factor is *semantic diversity* (Hoffman et al., 2013), which indicates the range of contexts in which a word can be used. Words with greater semantic diversity tend to have more meanings than those with lower diversity. This variability can enhance word recognition (Hoffman & Woollams, 2015), even when frequency, document count, and age of acquisition are accounted for. Words with high semantic diversity are more straightforward to process than those with low diversity, even after controlling for frequency (Pagán et al., 2020). The semantic diversity of a verbal prompt indicates the number of potential associative pathways available during ideation and can influence the diversity or uniformity of the responses. Theoretically, using words with lower semantic diversity will restrict the ideational pathways, leading to response similarity and lower originality. It is also possible that the semantic diversity of the prompt may influence whether the respondents utilize flexibility or persistence pathways more effectively (Nijstad et al., 2010). The flexibility pathway may be more often used with verbal prompts with a high semantic diversity, whereas the persistence pathway may be more important with verbal prompts with a lower semantic diversity.

Age of acquisition is another indicator showing the age at which a word is typically learned (Morrison & Ellis, 2000). Low-frequency words are typically

acquired later in life; therefore, there is a negative correlation between the age of acquisition and word frequency (Johnston & Barry, 2006). The age of acquisition of a word influences reaction times in word association and semantic categorization tasks (Brysbaert *et al.*, 2000; Catling & Johnston, 2005) and performance on lexical decision latency and naming tasks (Cortese & Khanna, 2007). Because the responses were produced by elementary-aged children (grades 3 through 5), the age of acquisition was included in the variable set.

Related to word frequency, we also included *bigram* (frequency of a letter pair in a word), item mean *reaction time* (mean reaction time to a word prompt), and *word length* (number of letters in a word) metrics in our model. Research shows that word frequency is negatively related to mean reaction time (Brysbaert *et al.*, 2018) and word length (Gerth & Festman, 2021). Essentially, words that are longer and or prompt a slower reaction time occur less frequently within language and by extension, could lead to restricted *prompt openness*. Similarly, bigrams serve as another metric of word frequency and have been shown to aid the accurate identification of low frequency words (Biederman, 1966; Chetail, 2015). Specifically, when words were shown tachistoscopically, low frequency words with low frequency bigrams in general had a lower threshold of word identification (Chetail, 2015), which could result in ease of recall and a correspondingly broader *prompt openness*. Lastly, prior research has found a significant interaction between time spent on task and frequency on participant fluency scores (Ogurlu *et al.*, 2023). When time on DT task was constrained word frequency negatively impacted fluency. Due to the interconnected nature of fluency and originality within DT assessments (Parnes, 1961) further examination is warranted for how these psycholinguistic indicators effect response originality.

The present study

Past research has indicated word frequency predicts fluency in Alternate Uses Test performance. In this present study, we explore the impact of specific psycholinguistic characteristics of *two different verbal divergent tests* (Alternate Uses and Instances) on the *originality* of the responses produced by elementary school children. Specifically, we analyzed the following psycholinguistic variables as independent variables: word frequency, semantic flexibility, word length, bigram, item reaction time, and age of acquisition. Because children responded to multiple DT prompts, we conducted multilevel analyses to account for dependency in the dataset where responses (Level 1) are nested within individuals (Level 2).

Methods

Sample

The study sample comprised 387 elementary-aged student participants who completed the Measure of Original Thinking in Elementary School (MOTES). Student participants were collected from six branches of the same public school network located in a major southwestern metropolitan area in the United States. Data were collected during school hours in groups ($n < 20$) using computers and

headsets. Groups were formed based on student schedules and parental consent. A single Qualtrics survey link, which included all MOTES prompts and instructions, was used to administer the assessment.

Approximately 386 third-, fourth-, and fifth-grade elementary school students ($Mage = 9.36$, $SD = .97$) completed the MOTES. Of the total sample, 137 (35.4%) were third graders, 132 (34.1%) fourth graders, and 118 (30.5%) fifth graders. The total sample consisted of 182 (47.1%) male and 205 (52.9%) female students. Concerning race and ethnicity, 125 (32.5%) were Hispanic or Latinx, 114 (29.6%) Black/African American, 60 (15.6%) White/European American, three (1%) American Indian/Indigenous identity, and eight with multiple ethnic identities (2.1%). 132 (34.1%) students were considered to be English language learners (ELLs), 60 (15.5%) were gifted and talented (GT) identified, and 49 (12.7%) received special education services. As compensation for participating, students received a toy gift.

Measure

The MOTES consists of 27 prompts across three subsequent MOTES tests: Test 1: How Do You Use It? (Alternate Uses); Test 2: What Is an Example? (Instances); and Test 3: Complete the Sentence. Each test consisted of nine items, one of which was a practice item to introduce participants to each test. For the purposes of this study, only responses from tests 1 and 2 will be utilized for further analysis because Test 3 involved prompts with multiple words, whereas Test 1 and Test 2 prompts had one root word. Additionally, one item from MOTES Test 1 (light bulb) was removed from further analysis because it is two words and thus did not have associated psycholinguistic metrics. With that in mind, this dataset's possible number of responses was 15×386 , or 5790. However, due to time limitations on children's responses, there were 450 missing responses (7.8%), for a total of 5340 responses, all of which were rated for originality using automated originality scoring.

Variables

The variance of response originality scores was distinguished at two levels to examine the effects of psycholinguistics on originality scores. Level 1 refers to the individual responses and their corresponding originality scores. Level 2 refers to the participants' overall originality performance. In essence, we seek to employ multilevel modeling to measure psycholinguistics' relative effect on response originality while simultaneously accounting for individual differences in originality.

All of the predictors tested were Level-1 variables. They refer to variables that directly relate to the prompts themselves and, by extension, the resulting originality of responses to said prompts. These variables consist of psycholinguistic metrics collected for all 15 of the included MOTES prompts. These metrics include SubtlWF (a standardized form of word frequency), word length, semantic diversity, age of acquisition, and bigram mean. SubtlWF, which indicates word frequency per million, was collected using archival data from Brysbaert and New (2009) and Brysbaert et al. (2018). Semantic diversity (SemD) scores were collected from Hoffman et al. (2013). SemD represents the log transformed and reversed sign of the

average cosine similarity between all the contexts of a given word (Hoffman et al., 2013). Therefore, words with a greater number of dissimilar contexts can be represented with a corresponding larger SemD. The age of acquisition (i.e., the average age in which a word is learned) data for included prompts were collected from Kuperman et al. (2012), including ratings for 30,121 English content words. Bigram mean, or the sum of two letter sequences contained within included prompts (e.g., “BALL”) divided by the number of successive two letter sequences (i.e., BA, AL & LL) collected by the English lexicon project (Balota et al., 2007) and included as another metric of word frequency. Word length was calculated by counting the number of letters within each of the included MOTES item prompts. In addition to psycholinguistic metrics, the type of question (i.e., DT prompt type) was included as a level 1 dummy coded variable with “Uses” prompts noted as 0 and “Instances” as 1. Finally, individual response originality served as the outcome variable at level 1 and participants average response originality served as the outcome variable at level 2. Originality scores were computed using automatic scoring following Acar et al. (2024) and Organisciak et al. (2023). All variables, except for *prompt type* were treated as continuous variables.

Procedure

Participants were administered the MOTES electronically using a single survey link to Qualtrics. To prepare text responses for scoring, all responses were reviewed by one of the lead researchers, who then removed double spaces and periods as well as indicated possible spelling errors. In order to best preserve student responses, no other edits were made to participant responses during preprocessing. Participants’ responses to MOTES were judged by a panel of five human raters and an automated originality scoring system. Raters were instructed to rate individually one prompt at a time on a 1 (*not original*) to 5 (*very original*) scale. They were instructed to follow a normal distribution for their ratings, with most scores being 2, 3, and 4 rather than extreme 1 or 5. As a result, raters show a high level of agreement ($\alpha = .844$). In the current study, we employed automated originality scoring through the Ocsai platform (Organisciak et al., 2023), which fine-tuned deep neural network-based large language models by training them on human-evaluated responses. Beyond scoring, large language models are also used for test item generation (Belzak et al., 2023). Our findings here can be used to extend the automation to the generation of test items by providing specificity around the nature of the items to select based on psycholinguistics. Automated scoring highly correlated with the human rating for How Do You Use It? ($r = .79$) and What Is an Example? ($r = .91$). Due to these high correlations and the recent shift in DT assessment scoring to automated scoring methods, only automated originality scores were used for further analysis.

Data analysis overview

The present study employs a 2-level hierarchical linear model (HLM) approach with individuals (i.e., student participants) at Level 2 and responses (i.e., ideas generated for MOTES Games 1 and 2) at Level 1. This allowed for the accounting of personal DT ability, which likely differs among study participants. It is also necessary because

of the possibility of variability between item responses by the same individual and between individuals. The details of the model and the study variables are presented below. Analyses were conducted using JASP 0.19.0.0 and SPSS Statistics 29.0.1.0. SPSS Statistics 29.0.1.0 was used to compute an unconditional model and test whether HLM was appropriate for data analysis. Subsequent models were analyzed using JASP 0.19.3.0.

First, an Unconditional Model (Model 1) was analyzed to test for differences in prompt automated originality at baseline (intercept) and students' average automated originality scores (slope). The model's residual variances were then used to calculate an intraclass correlation coefficient (ICC), quantifying automated originality variance at the item and student levels.

After testing Model 1, bivariate correlations between predictors were assessed for high degrees of multicollinearity. All predictors of interest were put into Model 2 to test their statistical significance simultaneously with associated random effects. Then, all predictors with non-statistically significant fixed effects (i.e., $p < .05$) were removed to form Model 3. Following that, Model 3 was assessed for its omnibus predictive capability compared to Model 1.

The retained predictors' random effects were then individually tested by examining the changes in model deviance relative to changes in model degrees of freedom with the assistance of a X^2 goodness of fit table. To aid in this process, Model 3 was rerun using restricted maximum likelihood to provide an appropriate deviance score to assess the effect of individual random effects being removed. This comparison model was then deemed Model 3a. Random effects were removed with replacement for each predictor until all retained predictors' random effects could be assessed. This procedure generated seven sub-models (Model 3b, 3c, 3d, 3e etc.) which were compared against Model 3a. The sub-model with the lowest model deviance relative to the comparison model, (i.e., Model 3a) was then selected as the new comparison model (Model 4a) and retained for further analysis. This was done to limit the possibility of nonsignificant random effects over inflating model deviance and masking the significance of other random effects. As a result, the single random effect with the largest decrease to model deviance associated with its removal was considered nonsignificant and therefore removed from further analysis. For the purpose of transparency and replicability, all associated data and relevant syntax have been posted to Open Science Framework (OSF) and can be reached using the following link <https://osf.io/e3gr8/>.

Results

The first step of the model building was the unconditional model (Model 1; see Table 1). Because the resultant ICC (.288) was greater than .10, it indicated the need to use multilevel modeling to account for data clustering (Lee, 2000). The unconditional model (Model 1) results suggest that the average student originality score is 2.875, with a starting residual variance of .267. Because several predictors were found to have medium to high correlation (i.e., $<|.4|$) (see Table 2), there was a concern for path dependency affecting the testing of subsequent models. In other words, the order in which predictors are added to subsequent models can drastically

Table 1. Psycholinguistic predictors and AI originality scores descriptive statistics

Variable	M	SD
Independent		
SubtlWF	101.971	168.376
Sem_D	1.617	.188
Bg_M	3100.580	1764.916
I_RT	577.901	44.883
AOA	3.597	.764
Length	4.60	1.819
Dependent		
Robo	2.875	.6112

Note: SubtlWF = standardized form of word frequency; Bg_M = bigram frequency mean; I_RT = item reaction time; AOA = age of acquisition; Robo = AI generated originality score.

Table 2. Predictor bivariate correlations

Predictor	SubtlWF	Sem_D	Bg_M	I_RT	AOA	Length	PT
SubtlWF	–	–	–	–	–	–	–
Sem_D	.635	–	–	–	–	–	–
Bg_M	–.133	.070	–	–	–	–	–
I_RT	–.461	–.592	.065	–	–	–	–
AOA	–.272	–.031*	.095	.378	–	–	–
Length	–.417	–.638	–.197	.585	.449	–	–
PT	.367	.683	–.051	–.551	.272	–.426	–

Note: *Indicates significance at $p < .05$ all other correlations are significant at $p < .001$ PT = Prompt Type.

affect the significance of individual predictors and their contribution to the omnibus model. Therefore, Model 2 was comprised of all test predictors (i.e., DT prompt type, age of acquisition, SubtlWF, bigram mean, item reaction time, word length, and semantic diversity) and three interaction terms, (Prompt Type*Semantic Diversity, Prompt Type* Word Length, and Prompt Type*Word Frequency). As shown in Tables 3–5, age of acquisition, word length, and bigram mean were found to be nonsignificant ($p > .05$). Item reaction time was also nonsignificant with $p = .067$. However, due to concerns of path dependency and inflated p -values from included random effects, item reaction time was retained while the other three predictors were removed. This resulted in Model 2b, which consisted of the remaining predictors, including item reaction time. As shown in Tables 3–5, all model 2b predictors were found to be statistically significant, $p < .05$ and were retained for further analysis.

Model 3a was generated using the same retained predictors from Model 2b and the restricted maximum likelihood (REML) function. As presented in Tables 3–5, Model 3a has an overall model deviance of 8593.992 with a corresponding 45 degrees of freedom. Successive models (3b–3h) were generated by removing individual retained predictors' random effects with replacement. Table 6 presents a complete report of all tested sub-models excluded random effects, resultant Δ deviances (Δ Dev), and Δ degrees of freedom (Δ df). Several Model 3 sub models did not show a significant increase in model deviance ($\chi^2 < 15.51$, $df = 8$) 3c (*prompt type* * *semantic diversity* random effect excluded Δ Dev = -0.577), 3d (*prompt type* random effect excluded Δ Dev = -9.956) and 3f (*semantic diversity* random effect excluded Δ Dev = 13.268).

To mitigate the possibility of Type II errors due to nonsignificant random effects inflating model deviance, only the random effect with the largest associated decrease in model deviance was deemed nonsignificant. All other random effects were retained for further analysis. Therefore, Model 3f (*semantic diversity* random effect excluded) was used as the new comparison model (i.e., Model 4a). Sub-models (4b–4g) were then generated following the same aforementioned procedure with model deviance, Δ Dev, and Δ df presented in Table 6. Sub-models 4f (*semantic diversity* & *prompt type***word length* random effect excluded Δ Dev = 83.679) and 4g (*semantic diversity* & *item reaction time* Δ Dev = 45.126) both displayed significant decreases in model deviance compared to model 4a. However, following the aforementioned precautions due to inflated model deviance, model 4f was selected as the new comparison model (i.e., Model 5a). Sub-models (i.e., 5b–5f) were generated by removing random effects with replacement and associated Δ Dev and Δ df reported in Table 6. Because all sub models showed significant increases in model deviance ($\chi^2 < 12.59$ $df = 6$), Model 5a was retained as the final model.

Therefore, Model 5a ($\chi^2 = 8497.045$ $df = 30$, $p < .05$) was retained as the final model with all included predictors' random effects included except for *semantic diversity* and *prompt type* by *word length*. To calculate the residual variance of the final model, we ran Model 5a using the maximum likelihood test method. The residual variance decreased significantly from the unconditional model (.267) to Model 5a (.218) (See Table 7). In other words, Model 5a explains approximately 18.4% of the variance in originality scores. That means 18.4% of the variability in the originality of the responses was explained by specific psycholinguistic characteristics of the selected prompts. According to this final model, response originality is positively related to semantic diversity ($b = .257$, $SE = .071$) and item reaction time ($b = .0005$, $SE = .0002$), whereas it is negatively related to word frequency ($b = -.002$, $SE = .0004$). Furthermore, response originality was more strongly associated with the Instances prompt than the Uses prompt ($b = .613$, $SE = .119$). Interaction effects of Prompt Type with Word Frequency ($b = .002$, $SE = .0004$), Semantic Diversity ($b = .506$, $SE = .134$), and Word length ($b = .057$, $SE = .012$) showed that response originality is larger when Instances prompts had greater word frequency, semantic diversity, or word length when compared to the Uses prompts. In other words, psycholinguistic characteristics seem to be more impactful on response originality for the Instances prompts than the Uses prompts.

Table 3. Model summaries

Parameters	Two-level empty model: model 1	All predictors added: model 2a	All significant fixed effects retained: model 2b	All significant fixed effects retained REML: Model 3a	PT*SWF random effect removed: model 3b	PT*Sem_D random effect removed: model 3c	PT random effect removed: model 3d	SWF random effect removed: model 3e	Sem_D random effect removed: model 3f
Fixed effects									
Intercept	2.875(.018)**	2.118(.148)**	2.139(.122)**	2.138(.123)**	2.136(.121)**	2.144(.127)**	2.142(.127)**	2.139(.121)**	2.139(.123)**
Bg_M_C	–	–.000005(.000004)	–	–	–	–	–	–	–
PT*SWF	–	0.002(.0004)**	.002(.0003)**	.002(.0003)**	.002(.0003)**	.002(.0004)**	.002(.0004)**	.002(.0003)**	.002(.0004)**
PT*SD	–	.558(.165)**	.510(.133)**	.511(.134)**	.512(.134)**	.506(.133)**	.507(.136)**	.512(.134)**	.512(.133)**
Prompt type	–	.651(.146)**	.617(.116)**	.618(.117)**	.621(.119)**	.611(.117)**	.613(.119)**	.618(.119)**	.617(.118)**
SWF_C	–	–.002(.0004)**	–.002(.0003)**	–.002(.0003)**	–.002(.0003)**	–.002(.0004)**	–.002(.0004)**	–.002(.0003)**	–.002(.0004)**
Sem_D_C	–	.259(.095)*	.258(.091)*	.258(.093)*	.26(.076)**	.258(.078)**	.255(.076)**	.256(.076)**	.255(.072)**
AoA_C	–	.0003(.026)	–	–	–	–	–	–	–
PT*Lg	–	.056(.013)**	.058(.013)**	.058(.013)**	.059(.016)**	.057(.012)**	.057(.014)**	.057(.015)**	.057(.013)**
Length_C	–	.002(.008)	–	–	–	–	–	–	–
I_RT_C	–	.0005(.0003)	.0005(.0002)*	.0005(.0002)*	.0005(.0003)*	.0005(.0003)*	.0005(.0003)*	.0005(.0003)*	.0005(.0002)*
Model summary									
Deviance	8828.839	8879.405	8510.125	8593.992	8705.310	8594.569	8603.948	8775.068	8580.724
AIC	8834.839	9035.405	8600.125	8683.992	8779.310	8668.569	8677.948	8849.068	8654.724
BIC	8854.588	9548.878	8896.359	8980.226	9022.881	8912.140	8921.519	9092.638	8898.295
Number of estimated parameters	3	78	45	45	37	37	37	37	37

Note: * $p < .05$. ** $p < .001$. Bg_M_C = mean centered *bigram frequency mean*; PT*SWF = interaction term between *prompt type* and *word frequency*, PT*SD = interaction term between *prompt type* and *semantic diversity*; SWF_C = mean centered *standardized form of word frequency*; Sem_D_C = mean centered *semantic diversity*; AOA_C = mean centered *age of acquisition*; PT*Lg = interaction term between *prompt type* and *word length*; I_RT_C = mean centered *item reaction time*.

Table 4. Model summaries

Parameters	PT*Lg random effect removed: model 3g	I_RT random effect removed: model 3h	Sem_D ran- dom effect removed: model 4a	Sem_D & PT*SWF random effects removed: model 4b	Sem_D & PT*SD random effects removed: model 4c	Sem_D & PT random effects removed model: 4d	Sem_D & SWF random effects removed: model 4e	Sem_D & PT*Lg random effects removed: model 4f	Sem_D & I_RT random effects removed: model 4g
Fixed effects									
Intercept	2.135(.125)**	2.130(.125)**	2.139(.123)**	2.139(.126)**	2.133(.124)**	2.144(.129)**	2.136(.121)**	2.139(.126)**	2.141(.125)**
Bg_M_C	–	–	–	–	–	–	–	–	–
PT*SWF	.002(.0003)**	.002(.0003)**	.002(.0004)**	.002(.0003)**	.002(.0004)**	.002(.0004)**	.002(.0003)**	.002(.0003)**	.002(.0003)**
PT*SD	.518(.133)**	.522(.136)**	.512(.133)**	.512(.136)**	.518(.131)**	.505(.134)**	.514(.134)**	.512(.136)**	.509(.136)**
Prompt type	.621(.118)**	.624(.119)**	.617(.118)**	.617(.121)**	.622(.116)**	.610(.118)**	.620(.119)**	.617(.121)**	.614(.120)**
SWF_C	–.002(.0003)**	–.002(.0003)**	–.002(.0004)*	–.002(.003)**	–.002(.0003)**	–.002(.0004)**	–.002(.0003)**	–.002(.0003)**	–.002(.0003)**
Sem_D_C	.256(.083)**	.250(.083)*	.255(.072)**	.256(.075)**	.256(.072)**	.256(.073)**	.257(.072)**	.256(.075)**	.255(.074)**
AoA_C	–	–	–	–	–	–	–	–	–
PT*Lg	.057(.012)**	.057(.013)**	.057(.013)**	.057(.012)**	.058(.017)**	.057(.013)**	.058(.015)**	.057(.012)**	.058(.013)**
Length_C	–	–	–	–	–	–	–	–	–
I_RT_C	.0005(.0002)*	.0005(.0002)*	.0005(.0002)*	.0005(.0002)*	.0005(.0002)*	.0005(.0003)*	.0005(.0003)	.0005(.0002)*	.0005(.0002)*
Model summary									
Deviance	8612.750	8642.481	8580.724	8616.702	9062.131	8660.163	8738.691	8497.045	8535.598
AIC	8686.750	8716.481	8654.724	8676.702	9122.131	8720.163	8798.691	8557.045	8595.598
BIC	8930.320	8960.051	8898.295	8874.191	9319.621	8917.653	8996.180	8754.534	8793.087
Number of estimated parameters	37	37	37	30	30	30	30	30	30

Note: * $p < .05$. ** $p < .001$. Bg_M_C = mean centered *bigram frequency mean*; PT*SWF = interaction term between *prompt type* and *word frequency*, PT*SD = interaction term between *prompt type* and *semantic diversity*; SWF_C = mean centered *standardized form of word frequency*; Sem_D_C = mean centered *semantic diversity*; AOA_C = mean centered *age of acquisition*; PT*Lg = interaction term between *prompt type* and *word length*; I_RT_C = mean centered *item reaction time*.

Table 5. Model summaries

Parameters	Sem_D & PT*LG random effects removed: model 5a	Sem_D, PT*LG, & PT*SWF random effects removed: model 5b	Sem_D, PT*LG, & PT*SD random effects removed: model 5c	Sem_D, PT*LG, & PT random effects removed: model 5d	Sem_D, PT*LG, & SWF random effects removed: model 5e	Sem_D, PT*LG, & I_RT random effects removed: model 5f	Sem_D, PT*LG random effects removed ML: final model
Fixed effects							
Intercept	2.139(.126)**	2.141(.122)**	2.129(.133)**	2.130(.125)**	2.140(.123)**	2.142(.122)**	2.144(.124)**
Bg_M_C	–	–	–	–	–	–	–
PT*SWF	.002(.0003)**	.002(.0003)**	.002(.0004)**	.002(.0003)**	.002(.0003)**	.002(.0003)**	.002(.0003)**
PT*SD	.512(.136)**	.510(.133)**	.523(.137)**	.522(.135)**	.511(.133)**	.509(.132)**	.506(.134)**
Prompt type	.617(.121)**	.616(.119)**	.625(.121)**	.625(.118)**	.617(.119)**	.615(.120)**	.613(.119)**
SWF_C	–.002(.0003)**	–.002(.0003)**	–.002(.0004)*	–.002(.0003)**	–.002(.0003)**	–.002(.0003)**	–.002(.0003)**
Sem_D_C	.256(.075)**	.259(.083)*	.255(.075)**	.256(.073)**	.257(.073)**	.259(.072)**	.257(.073)**
AoA_C	–	–	–	–	–	–	–
PT*Lg	.057(.012)**	.057(.013)**	.057(.012)**	.057(.012)**	.057(.012)**	.058(.012)**	.057(.012)**
Length_C	–	–	–	–	–	–	–
I_RT_C	.0005(.0002)*	.0005(.0002)*	.0005(.0002)*	.0005(.0002)**	.0005(.0002)**	.0005(.0002)*	.0005(.0002)*
Model summary							
Deviance	8497.045	8763.554	8685.871	8590.049	8715.204	8883.677	8492.581
AIC	8557.045	8811.554	8733.871	8638.049	8763.204	8931.677	8552.581
BIC	8754.534	8898.295	8891.862	8796.040	8921.196	9089.669	8750.071
Number of estimated parameters	30	24	24	24	24	24	30

Note: * $p < .05$. ** $p < .001$. Bg_M_C = mean centered *bigram frequency mean*; PT*SWF = interaction term between *prompt type* and *word frequency*, PT*SD = interaction term between *prompt type* and *semantic diversity*; SWF_C = mean centered *standardized form of word frequency*; Sem_D_C = mean centered *semantic diversity*; AOA_C = mean centered *age of acquisition*; PT*LG = interaction term between *prompt type* and *word length*; I_RT_C = mean centered *item reaction time*.

Table 6. Model deviance for excluded random effects

Excluded Random Effect	Deviance	df	Δ Dev	Δ df
Model 3a REML				
None	8593.992	45	–	–
PT*SubtlWf	8705.310	37	–111.318	8
PT*Sem_D	8594.569	37	–.577	8
PT	8603.948	37	–9.956	8
SubtlWf	8775.068	37	–181.076	8
Sem_D	8580.724	37	13.268	8
PT*Lg	8612.750	37	–18.758	8
I_RT	8642.481	37	–48.489	8
Model 4a REML				
Sem_D	8580.724	37	–	–
Sem_D & PT*SWF	8616.702	30	–35.978	7
Sem_D & PT*Sem_D	9062.131	30	–481.407	7
Sem_D & PT	8660.163	30	–79.439	7
Sem_D & SubtlWf	8738.691	30	–157.967	7
Sem_D & PT*Lg	8497.045	30	83.679	7
Sem_D & I_RT	8535.598	30	45.126	7
Model 5a REML				
Sem_D & PT*Lg	8497.045	30	–	–
Sem_D, PT*Lg, & PT*SWF	8763.554	24	–266.509	6
Sem_D, PT*Lg, & PT*Sem_D	8685.871	24	–188.826	6
Sem_D, PT*Lg, & PT	8590.049	24	–93.004	6
Sem_D, PT*Lg, & SubtlWf	8715.204	24	–218.159	6
Sem_D, PT*Lg, & I_RT	8883.677	24	–386.632	6

Discussion

In the present study, we tested the relationship between response originality and psycholinguistic characteristics of the DT prompts. Our multilevel model showed that psycholinguistics characteristics could explain about 23% of the variability in original thinking. Specifically, word frequency, semantic diversity, reaction time, and word length were significant predictors. Furthermore, DT prompt type (Instances versus Uses) moderated their influence.

Word frequency

We found that word frequency is *negatively* related to response originality. In fact, word frequency was the only negative predictor in the model. That means responses

Table 7. Model comparison of fixed effects and residual variances

Unconditional				
Term	Estimate	<i>t</i>	<i>p</i>	Variance
Model 4a				
Term	Estimate	<i>t</i>	<i>p</i>	–
Intercept	2.875	158.713	<.001	–
Residual	–	–	–	.267
Intercept	2.144	17.261	<.001	–
PT*SubtlWF	.002	6.911	<.001	–
PT*Sem_D	.506	3.780	<.001	–
Prompt Type	.613	5.127	<.001	–
SubtlWF_C	–.002	–6.334	<.001	–
Sem_D_C	.257	3.509	<.001	–
PT*Length	.057	4.771	<.001	–
I_RT_C	.0005	2.380	.018	–
Residual	–	–	–	.218

Note: C = refers to the listed variable being mean centered. To aid in interpretation of predictor fixed effects, SubtlWF, Sem_D, and I_RT were centered by their means.

were less original when DT prompts were picked from those with a larger word frequency value. Larger word frequency implies greater accessibility and lower path resistance toward the prompt. It may be that experienced, or what some call “old ideas” (Benedek et al., 2014a), predominates idea generation process, leaving less room for “new ideas,” which are employed when generating responses that are not directly tied to past experiences (Gilhooly et al., 2007). These findings partly conflict with Mednick’s (1962) associative theory of creativity, asserting that greater associations increase the probability of finding a creative solution. Here, we observed the opposite effect, where responding to items with greater Word frequency, which are expected to elicit more associations, is associated with lower originality. Importantly, Mednick (1962) did not manipulate the prompts but investigated individual differences in associative ability. He compared the association patterns between high- and low-creative individuals and proposed highly creative ones to have a flat associative hierarchy, characterized by a smaller slope of production over time compared to the low-creative individuals who quickly run out of associations after generating a lot early on.

On the other hand, it is also possible to interpret our findings in line with Mednick’s (1962) perspective that word frequency allows for quick idea generation followed by a dramatic decline (increasing greater negative slope) and therefore resulting in diminished chance of getting at an original response. Notably, past research showed that word frequency is related to more idea quantities in the Alternate Uses test when people are given a short time (Forthmann et al., 2016; Ogurlu et al., 2023). So, the link between word frequency and response quantity is

well established. However, generating more responses is not necessarily conducive to having original ideas. In fact, the opposite appears to be true. Gonthier and Besançon (2024) found that fluency negatively affects originality and elaboration: individuals producing fewer ideas had higher mean originality and elaboration scores. In a recent study, time use appeared to explain this: greater time use (latency) was positively related to originality and negatively related to fluency (Acar et al., 2019, n.d.).

Reaction time, word length, and age of acquisition

Reaction time was positively related to originality. This finding reinforces our interpretation of the negative relationship between originality and word frequency. Words that trigger slower reactions (longer latency) tend to have a lower word frequency (Johnston & Barry, 2006). Consequently, picking words with a longer reaction time may result in less automated, more deliberate, purposeful, and effortful cognitive processes. This is consistent with recent research that underlines the critical role of executive functions in creative processes (Benedek et al., 2014; Zabelina et al., 2019).

On the other hand, age of acquisition and word length did not significantly affect originality. Research shows that age of acquisition and word length are negatively related to word frequency (Johnston & Barry, 2006). It may be that their presence in the same model, along with word frequency, has led to a masking effect, where the potential influence of age of acquisition and word length has been masked or accounted for by word frequency. However, that does not mean word length is irrelevant. We found a significant interaction between word length and prompt type, showing that the impact of word length is specific to DT prompt type. This is discussed later in the Discussion section.

Semantic diversity

We found a positive relationship between semantic diversity and originality. Semantic diversity measures how much a word's meaning changes or differs across various situations, contexts, or uses (Hoffman et al., 2013). A word with high semantic diversity is used in a wide range of contexts with varying interpretations. In contrast, a word with low semantic diversity is used in more consistent or similar contexts. Semantic diversity indicates the degree to which a word can be used flexibly due to its ambiguity. Thus, ambiguous words tend to elicit more diverse meanings and associations, increasing various interpretations and diminishing the possibility of one following a standard ideational pattern. In turn, words with a high semantic diversity are expected to contribute to original thinking. This very definition of semantic diversity aligns well with the definition of DT: the ability to think in various directions (Runco, 1999). As a result, it appears that picking words with a high level of semantic diversity could serve well as a DT prompt by eliciting more original responses on average.

Importantly, semantic diversity did not contribute to the quantity of responses (Ogurlu et al., 2023). The fact that it contributed to originality but not quantity can be interpreted in terms of the dual pathway theory of creativity (Nijstad et al., 2010). According to this theory, creativity is driven by two processes: persistence, the ability

to continue working on a problem despite challenges, and flexibility, the ability to switch between different ideas or strategies. These processes work together to generate novel solutions, with persistence ensuring sustained effort and flexibility, allowing for the exploration of diverse possibilities. It may be that greater semantic diversity is conducive to utilizing the flexibility pathway, which is more supportive of original ideation. Conversely, word frequency is more strongly linked to the persistence path, which is more suitable for greater productivity. Because the MOTES is a single-response test, a direct test of the relationship among semantic diversity, word frequency, and ideational fluency was not testable in the present work.

Moderating influence of prompt type

Interestingly, the relationship between word frequency and originality was moderated by prompt type. That means, as word frequency increases, the negative relationship between word frequency and originality becomes weaker for the Instances prompts when compared to Uses. We also found a moderating effect of prompt type for semantic diversity and word length. The relationship between semantic diversity and response originality was greater (i.e., more strongly positive) for Instances than Uses. Likewise, the relationship between word length and response originality was greater for Instances than Uses. Note that word length did not have a significant main effect. Thus, its influence on the originality of responses is limited to the Instances task. Research has shown that word length is related to recall success: shorter words are more easily recalled than longer words (Baddeley, 1999). Longer words are more likely to be avoided in everyday language because they are considered uneconomic (Zipf, 1965[1935]). Neuroscientific research also showed that longer words trigger higher neurophysiological activity in the early phases, which may have been caused by lower familiarity with them (Hauk & Pulvermüller, 2004).

The consistency of the moderating influence of prompt type across semantic diversity, word frequency, and word length requires a close look into the differences between the Uses and Instances tasks. The Instances task asks respondents to identify things that belong to a specific category (i.e., things that are “red” or “big”), whereas Uses require finding uses for a common object. Silvia (2011) compared these two tasks and argued that the Uses task is more likely to engage executive functions, whereas the Instances task is more associational. This also explains the inapplicability of Mednick’s (1962) hypothesis to the Uses task. Mednick’s work was based on the word association’s tasks, which more closely resemble Instances prompts over the Uses test.

Implications

The findings of this study underscore that performance on originality in verbal DT tests depends to some degree on the psycholinguistic characteristics of the word prompts used. These findings have important implications for both theory and measurement. From a theoretical perspective, the results suggest that linguistic features shape the cognitive accessibility of associative networks that underlie idea generation leading to greater or lower originality. Consequently, creative

performance on verbal DT tasks is not solely a function of an individual's creative potential but is also partially determined by the linguistic affordances of the stimulus itself (Acar & Runco, 2024; Runco, 1992).

From a measurement standpoint, these findings emphasize that the arbitrary selection of prompts can introduce systematic variance unrelated to the construct being measured, potentially biasing interpretations of originality scores. This is particularly relevant in pre- and post-designs, where differences between test administrations could reflect item-level differences rather than true change in creative ability. To improve fairness and validity, test developers should therefore standardize prompt selection based on psycholinguistic equivalence rather than surface content similarity. On the other hand, it is important to keep in mind the target construct being measured through DT tasks. Our findings, along with past evidence (Ogurlu et al., 2023), indicated that what benefits originality appears to work against fluency as in the case of word frequency. Therefore, the scorable outcomes of DT tasks must be determined in advance to not only inform the explicit instructions (Reiter-Palmon et al., 2019) but also the type and use of psycholinguistics for prompt selection. Our findings are also directly applicable to the development of parallel test forms. When constructing alternate forms of verbal DT assessments, selecting prompts that match on key psycholinguistic parameters such as word frequency and semantic diversity can minimize unintended difficulty differences and ensure comparable measurement across administrations. Such alignment would enhance measurement invariance (Embretson & Reise, 2000) and facilitate more accurate interpretations of pre-post or longitudinal changes in creative performance.

Furthermore, our findings extend to AI-based creativity assessment. Recent advances, such as the Ocsai platform (Organisciak et al., 2023), employ large language models fine-tuned on human-rated DT responses to automate originality scoring. MOTES is scored using the methods provided in Ocsai. Beyond scoring, large language models have also been used for automated item generation (Belzak et al., 2023). The patterns observed here can inform such algorithms by identifying which linguistic characteristics yield balanced yet discriminating DT prompts. Currently, MOTES has a validated set of items but could use a dynamic item pool to draw from. The present results provide empirical guidance for refining and improving these systems: incorporating psycholinguistic indicators as item selection or weighting criteria could improve the validity and interpretability of automated scoring outputs.

Collectively, these implications converge on a central insight: psycholinguistic calibration is integral to the valid assessment of creative thinking. By accounting for linguistic properties in both human- and AI-mediated testing contexts, researchers and practitioners can reduce construct-irrelevant variance, improve fairness, and enhance the replicability and interpretability of findings across studies and populations (Messick, 1995).

Limitations and future directions

Several limitations of the present study should be acknowledged, some of which suggest promising directions for future research. First, participants were asked to produce only a single response per prompt. It remains unclear whether similar

psycholinguistic effects on originality would emerge under alternative task structures that allow multiple responses (e.g., three or unlimited). Allowing participants to generate additional ideas may engage broader associative networks and alter the influence of linguistic variables such as word frequency. Prior research has shown that later responses in DT tasks tend to be more original, a pattern known as the serial order effect (Beaty & Silvia, 2012; Christensen et al., 1957). Because of the design of MOTES, this effect could not be examined in the current study. Future research should replicate the present findings using DT task structure that permits multiple responses to determine whether psycholinguistic influences remain consistent across response positions. Moreover, future studies may explore whether the psycholinguistic characteristics of DT prompts themselves modulate the magnitude or direction of the serial order effect.

Second, the study sample consisted exclusively of elementary school children, which may constrain the generalizability of the findings. Psycholinguistic properties may exert stronger effects in this population because children's vocabularies and semantic networks are still developing (Wulff et al., 2022). It is therefore essential to examine whether these patterns hold in older youth and adults, whose more extensive lexical knowledge and metacognitive strategies may reduce sensitivity to linguistic affordances (Cosgrove et al., 2023).

Third, potential time-related effects represent an important consideration for future research. In the MOTES administration, participants were provided a sufficient but finite period of time to generate their responses. However, differences in allotted time could substantially influence how psycholinguistic variables affect performance. Ogurlu et al. (2023) found that word frequency, which tends to enhance fluency, has a stronger effect when task duration is short. Extending or shortening the response window could therefore shift the balance fluency and originality. Future studies should systematically manipulate time constraints to clarify how temporal conditions moderate psycholinguistic effects.

Building on these methodological considerations, future research should also extend this line of inquiry to other DT tasks (e.g., Similarities) and other measures of verbal creativity to determine whether the observed psycholinguistic effects generalize across task types. While the current study focuses on originality, subsequent work should examine flexibility and appropriateness as complementary indicators of creative performance. Further, examining interactions between psycholinguistic characteristics and demographic variables (i.e., race, ethnicity, and English language learner status) would help identify potential sources of differential item functioning and enhance the fairness of creativity assessments.

Replication package. All associated data and relevant syntax have been posted to Open Science Framework (OSF) and can be reached using the following link <https://osf.io/e3gr8>.

References

- Acar, S., Abdulla, A. M., & Runco, M. A. (2019). Latency as a predictor of originality in divergent thinking. *Thinking Skills and Creativity*, 33, 10574. <https://doi.org/10.1016/j.tsc.2019.100574>
- Acar, S., Berthiaume, K. C., Kagan, S., Dumas, D., & Organisciak, P. (n.d.) In praise of slow idea generation: Latency differentially predicts ideational quantity and quality. Manuscript in progress.

- Acar, S., Burnett, C., & Cabra, J. F. (2017). Ingredients of creativity: Originality and more. *Creativity Research Journal*, 29(2), 133–144. <https://doi.org/10.1080/10400419.2017.1302776>
- Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology*, 116(6), 953–981. <https://doi.org/10.1037/edu0000844>
- Acar, S., & Runco, M. A. (2024). Creativity assessment: Reactions, reflections, and directions for future research. In M. A. Runco & S. Acar (Eds.), *Handbook of Creativity Assessment* (pp. 335–361). Edward Elgar Publishing.
- 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>
- Baddeley, A. (1999). *Essentials of human memory*. Psychology Press.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445–459. <https://doi.org/10.3758/BF03193014>
- Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, 6(4), 309–319. <https://doi.org/10.1037/a0029171>
- Belzak, W. C., Naismith, B., & Burstein, J. (2023, June). Ensuring fairness of human-and AI-generated test items. *International Conference on Artificial Intelligence in Education*. Cham: Springer Nature Switzerland (pp. 701–707).
- Benedek, M., Jauk, E., Fink, A., Koschutnig, K., Reishofer, G., Ebner, F., & Neubauer, A. C. (2014a). 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., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014b). Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence*, 46, 73–83. <https://doi.org/10.1016/j.intell.2014.05.007>
- Biederman, G. B. (1966). Supplementary report: The recognition of tachistoscopically presented five-letter words as a function of digram frequency. *Journal of Verbal Learning and Verbal Behavior*, 5(2), 208–209. [https://doi.org/10.1016/S0022-5371\(66\)80020-8](https://doi.org/10.1016/S0022-5371(66)80020-8)
- British National Corpus Consortium. (2007). British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. <https://www.natcorp.ox.ac.uk/>
- Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1), 45–50. <https://doi.org/10.1177/0963721417727521>
- Brysbaert, M., & New, B. (2009). Subtlexus: American word frequencies. http://www.lexique.org/?page_id=241.
- Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104, 215–226. [https://doi.org/10.1016/S0001-6918\(00\)00021-4](https://doi.org/10.1016/S0001-6918(00)00021-4)
- Catling, J. C., & Johnston, R. A. (2005). Age of acquisition effects in word generation. *European Journal of Cognitive Psychology*, 17, 161.
- Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology*, 6, 645. <https://doi.org/10.3389/fpsyg.2015.00645>
- Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology*, 53(2), 82–88. <https://doi.org/10.1037/h0045461>
- Clark, P. M., & Mirels, H. L. (1970). Fluency as a pervasive element in the measurement of creativity. *Journal of Educational Measurement*, 7(2), 83–86. <https://doi.org/10.1111/j.1745-3984.1970.tb00699.x>
- Cofer, C. N., & Shevitz, R. (1952). Word-association as a function of word-frequency. *The American Journal of Psychology*, 65(1), 75–79.
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, 60, 1072–1082. <https://doi.org/10.1080/17470210701315467>

- Cosgrove, A. L., Beaty, R. E., Diaz, M. T., & Kenett, Y. N. (2023). Age differences in semantic network structure: Acquiring knowledge shapes semantic memory. *Psychology and Aging*, *38*(2), 87–102. <https://doi.org/10.1037/pag0000721>
- Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391–404. https://doi.org/10.1207/s15326934crj1803_13
- Davies, M. (2008). The corpus of contemporary American English (COCA): 520 million words, 1990-present. Brigham Young University. <https://www.english-corpora.org/coca/>.
- Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*(1), 35–40. <https://doi.org/10.1037/a0038688>
- Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(4), 645. <https://doi.org/10.1037/aca0000319>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition*. MIT Press.
- Forster, E. (2008). *The role of function, homogeneity and syntax in creative performance on the uses of objects task*. University of Toronto [Unpublished master's thesis].
- Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32. <https://doi.org/10.1016/j.intell.2016.03.005>
- Gerth, S., & Festman, J. (2021). Reading development, word length and frequency effects: An eye-tracking study with slow and fast readers. *Frontiers in Communication*, *6*, 743113. <https://doi.org/10.3389/fcomm.2021.743113>
- Gerver, C. R., Griffin, J. W., Dennis, N. A., & Beaty, R. E. (2023). Memory and creativity: A meta-analytic examination of the relationship between memory systems and creative cognition. *Psychonomic Bulletin & Review*, *30*(6), 2116–2154. <https://doi.org/10.3758/s13423-023-02303-4>
- Gerwig, A., Miroshnik, K., Forthmann, B., Benedek, M., Karwowski, M., & Holling, H. (2021). The relationship between intelligence and divergent thinking—A meta-analytic update. *Journal of Intelligence*, *9*(2), 23. <https://doi.org/10.3390/jintelligence9020023>
- Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98*(4), 611–625. <https://doi.org/10.1111/j.2044-8295.2007.tb00467.x>
- Gonthier, C., & Besançon, M. (2024). It is not always better to have more ideas: Serial order and the trade-off between fluency and elaboration in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *18*(4), 480–492. <https://doi.org/10.1037/aca0000485>
- Guilford, J. P., Christensen, P. R., & Merrifield, P. R. (1958). *Consequences: Manual for administration, scoring, and interpretation*. Sheridan Psychological Services.
- Guilford, J. P., Merrifield, P. R., & Wilson, R. C. (1958). *Unusual uses test*. Sheridan Psychological Services.
- Harley, T. A. (2013). *The psychology of language: From data to theory*. Psychology press.
- Hass, R. W., & Beaty, R. E. (2018). Use or consequences: Probing the cognitive difference between two measures of divergent thinking. *Frontiers in Psychology*, *9*, 2327. <https://doi.org/10.3389/fpsyg.2018.02327>
- Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103. <https://doi.org/10.1016/j.clinph.2003.12.020>
- Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*, 718–730. <https://doi.org/10.3758/s13428-012-0278-x>
- Hoffman, P., & Woolams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 385–402. <https://doi.org/10.1037/a0038995>
- Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*(7–8), 789–845. <https://doi.org/10.1080/13506280544000066>
- Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. John Wiley & Sons.
- Kaya, F., & Acar, S. (2019). The impact of originality instructions on cognitive strategy use in divergent thinking. *Thinking Skills and Creativity*, *33*, 100581. <https://doi.org/10.1016/j.tsc.2019.100581>

- Kim, K. H. (2008). Meta-analyses of the relationship of creative achievement to both IQ and divergent thinking test scores. *The Journal of Creative Behavior*, 42(2), 106–130. <https://doi.org/10.1002/j.2162-6057.2008.tb01290.x>
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990.
- Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, 35(2), 125–141. https://doi.org/10.1207/S15326985EP3502_6
- Madore, K. P., Addis, D. R., & Schacter, D. L. (2015). Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. *Psychological Science*, 26(9), 1461–1468. <https://doi.org/10.1177/0956797615591863>
- Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3), 220–232. <https://doi.org/10.1037/h0048850>
- Menn, L., & Dronkers, N. F. (2016). *Psycholinguistics: Introduction and applications*. Plural publishing.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Miroshnik, K. G., Forthmann, B., Karwowski, M., & Benedek, M. (2023). The relationship of divergent thinking with broad retrieval ability and processing speed: A meta-analysis. *Intelligence*, 98, 101739. <https://doi.org/10.1016/j.intell.2023.101739>
- Monzell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118, 43–71.
- Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, 91(2), 167–180. <https://doi.org/10.1348/000712600161763>
- Nijstad, B. A., De Dreu, C. K., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>
- Ogurlu, U., Acar, S., & Ozbey, A. (2023). Does word frequency impact ideational fluency in divergent thinking? A meta-analytic exploration with the alternate uses test. *Thinking Skills and Creativity*, 47, 101139. <https://doi.org/10.1016/j.tsc.2022.101139>
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49, 101356. <https://doi.org/10.1016/j.tsc.2023.101356>
- Paek, S. H., Alabbasi, A. M. 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*, 41, 100894. <https://doi.org/10.1016/j.tsc.2021.100894>
- Pagán, A., Bird, M., Hsiao, Y., & Nation, K. (2020). Both semantic diversity and frequency influence children's sentence reading. *Scientific Studies of Reading*, 24(4), 356–364. <https://doi.org/10.1080/10888438.2019.1670664>
- Parnes, S. J. (1961). Effects of extended effort in creative problem solving. *Journal of Educational Psychology*, 52(3), 117–122. <https://doi.org/10.1037/h0044650>
- Plucker, J. A., Meyer, M. S., & Liu, P. (2024). Divergent thinking: Early views. In M. A. Runco, & S. Acar (Eds.). *Handbook of Creativity Assessment* (pp. 114–127). Edward Elgar.
- 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>
- Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2014). Effects of problem scope and creativity instructions on idea generation and selection. *Creativity Research Journal*, 26(2), 185–191. <https://doi.org/10.1080/10400419.2014.901084>
- Rothenberg, A., & Hausman, C. R. (1976). *The creativity question*. Duke University Press.
- Runco, M. A. (1988). Creativity research: Originality, utility, and integration. *Creativity Research Journal*, 1(1), 1–7. <https://doi.org/10.1080/10400418809534283>
- Runco, M. A. (1992). Children's divergent thinking and creative ideation. *Developmental Review*, 12(3), 233–264.
- Runco, M. A. (1999). Divergent Thinking. In M. A. Runco, & S. R. Pritzker (Eds.), *Encyclopedia of Creativity* (1, pp. 577–82). Academic Press.

- Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best. *Creativity. Theories–Research–Applications*, 3(1), 4–18. <https://doi.org/10.1515/ctra-2016-0001>
- Runco, M. A., & Acar, S. (2010). Do tests of divergent thinking have an experiential bias? *Psychology of Aesthetics, Creativity, and the Arts*, 4(3), 144–148. <https://doi.org/10.1037/a0018969>
- Runco, M. A., & Acar, S. (2024a). Assessing Divergent Thinking: Recent Advances. In M. A. Runco, & S. Acar (Eds.), *Handbook of creativity assessment* (pp. 128–147). Edward Elgar.
- Runco, M. A., & Acar, S. (Eds.). (2024b). *Handbook of creativity assessment*. Edward Elgar Publishing.
- Runco, M. A., & Albert, R. S. (1985). The reliability and validity of ideational originality in the divergent thinking of academically gifted and nongifted children. *Educational and Psychological Measurement*, 45(3), 483–501. <https://doi.org/10.1177/001316448504500306>
- Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *Journal of Creative Behavior*, 39(2), 137–148. <https://doi.org/10.1002/j.2162-6057.2005.tb01255.x>
- Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96. <https://doi.org/10.1080/10400419.2012.650092>
- Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, 6(1), 24–30. <https://doi.org/10.1016/j.tsc.2010.06.001>
- Silvia, P. J., Beaty, R. E., & Nusbaum, E. C. (2013). Verbal fluency and creativity: General and specific contributions of broad retrieval ability (Gr) factors to divergent thinking. *Intelligence*, 41(5), 328–340. <https://doi.org/10.1016/j.intell.2013.05.004>
- Stein, M. I. (1953). Creativity and culture. *The Journal of Psychology*, 36(2), 311–322. <https://doi.org/10.1080/00223980.1953.9712897>
- Vitrano, D., Altarriba, J., & Leblebici-Basar, D. (2021). Revisiting Mednick's (1962) theory of creativity with a composite measure of creativity: The effect of stimulus type on word association production. *The Journal of Creative Behavior*, 55(4), 925–936. <https://doi.org/10.1002/jocb.498>
- Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children*. Holt, Rinehart and Winston.
- 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>
- Witczak, O., Krzysik, I., Bromberek-Dyzman, K., Thierry, G., & Jończyk, R. (2024). Controlling stimulus ambiguity reduces spurious creative ideation variance in a cyclic adaptation of the alternative uses task. *Scientific Reports*, 14(1), 12492. <https://doi.org/10.1038/s41598-024-63225-2>
- Wulff, D. U., Hills, T. T., & Mata, R. (2022). Structural differences in the semantic networks of younger and older adults. *Scientific Reports*, 12(1), 21459. <https://doi.org/10.1038/s41598-022-11698-4>
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46, 441–517. <https://doi.org/10.1006/jmla.2002.2864>
- Zabelina, D. L., Friedman, N. P., & Andrews-Hanna, J. (2019). Unity and diversity of executive functions in creativity. *Consciousness and Cognition*, 68, 47–56. <https://doi.org/10.1016/j.concog.2018.12.005>
- Zipf, G. K. (1965 [1935]). *The psycho-biology of language*. MIT Press.