Creative or Not? Hierarchical Diffusion Modeling of the Creative Evaluation Process

Michelle C. Donzallaz[1], Julia M. Haaf[1], & Claire E. Stevenson[1]

[1] University of Amsterdam

Draft version 2, May 2022. This paper has not been peer reviewed. Please do not copy or cite without authors' permission.

Author Note

This report was written in R-Markdown with code for data analysis integrated into the text. The source code and the data can be found at: https://osf.io/73c2d/. This research was presented at the virtual MathPsych/ICCM conference in 2020 and was funded by the Amsterdam Brain & Cognition (ABC) Talent Grant (University of Amsterdam) 2016-2018 awarded to CES and the Jacobs Foundation Fellowship 2019-2022 awarded to CES (2018 1288 12). MCD was supported by a Vidi grant (VI.Vidi.191.091 to D. Matzke) from the Netherlands Organization of Scientific Research (NWO). JMH was supported by a Veni grant from the NWO (VI.Veni.201G.019). We thank Henrik Singmann and Han van der Maas for their valuable feedback on earlier versions of this manuscript.

Correspondence concerning this article should be addressed to Michelle C. Donzallaz, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam. E-mail: m.c.donzallaz@uva.nl

Abstract

When producing creative ideas (i.e., ideas that are original and useful) two main processes occur: ideation, where people brainstorm ideas, and evaluation, where they decide if the ideas are creative or not. While much is known about the ideation phase, the cognitive processes involved in creativity evaluation are less clear. In this paper, we present a novel modeling approach for the evaluation phase of creativity. We apply the drift diffusion model (DDM) to the creative-or-not (CON)-task to study the cognitive basis of evaluation and to examine individual differences in the extent to which people take originality and utility into account when evaluating creative ideas. The CON-task is a timed decision-making task where participants indicate whether they find uses for certain objects creative or not (e.g., using a book as a buoy). The different uses vary on the two creativity dimensions 'originality' and 'utility'. In two studies (n = 293, 17806 trials; n = 152, 9291 trials), we found that stimulus originality was strongly related to participants' drift rates but found only weak evidence for an association between stimulus utility and the drift rate. However, participants differed substantially in the effects of originality and utility. Furthermore, the implicit weights assigned to originality and utility on the CON-task were aligned with self-reported importance ratings of originality and utility and associated with divergent thinking performance in the alternative uses task (AUT). This research provides a cognitive modeling approach to creativity evaluation and underlines the importance of communicating rating criteria in divergent thinking tasks to ensure a fair assessment of creative ability.

Keywords: creativity, evaluation, diffusion model, Bayesian hierarchical modeling

Creative or Not? Hierarchical Diffusion Modeling of the Creative Evaluation Process

Creative ideas are essential for tackling today's problems, from personal obstacles, such as combining work and childcare during a pandemic lockdown, to societal challenges such as climate change (Hennessey & Amabile, 2010). It is no wonder that educators are increasingly including creativity in curricula and that managers consider creativity a key skill (Casner-Lotto & Barrington, 2006; IBM, 2010). Divergent thinking tasks are often used to assess creative ability, the ability to produce original and useful ideas (Barron, 1955; Runco & Jaeger, 2012; Stein, 1953). Perhaps the most common divergent thinking measure is the alternative uses task (AUT; Guilford, 1967; Runco & Acar, 2012). On the AUT, people are typically asked to come up with as many unusual uses as possible for a given object (e.g., "bath toy" for the object "brick") within a certain time interval. When solving the AUT, two main processes occur: ideation and evaluation (Basadur, 1995; Guilford, 1967; Runco & Acar, 2012). Ideation is the "brainstorm" phase where one comes up with ideas, and evaluation is the decision making phase where one judges which ideas are creative enough to pursue (or in the case of the AUT to list as a response). While most research has focused on ideation and how to improve it (e.g., Benedek, Fink, & Neubauer, 2006; Forthmann et al., 2019), how people evaluate and select ideas is less understood (Grohman, Wodniecka, & Kłusak, 2006; Ritter, van Baaren, & Dijksterhuis, 2012; Silvia, 2008). Yet, in the real world, it only takes one well-selected idea to solve a problem (e.g., the printing press reduced the cost and labor of printing books, thereby increasing literacy and making knowledge more accessible) or disregarding a good idea to lose the battle in innovative business (e.g., Blockbuster vs. Netflix; Randolph, 2019). In this paper, we take a cognitive modeling approach to fill this gap and study how people decide which ideas are creative or not on the AUT.

While the standard definition of creativity states that originality and utility are both needed for creativity, individuals may differ in how much they value these two dimensions. Previous research has mainly focused on population-level effects and found that people tend

to value originality more than utility when judging the creativity of an idea (Caroff & Besançon, 2008; Diedrich, Benedek, Jauk, & Neubauer, 2015; Runco & Charles, 1993). However, creativity is a widely used word and regularly discussed by laypeople (Davies, 2008; Mueller, Melwani, Loewenstein, & Deal, 2018). Consequently, different people may have substantially different conceptions of how important originality and utility are for creativity (e.g., Loewenstein & Mueller, 2016). For example, some may value utility more than others when judging creativity, and some may even find ideas or products that are not useful more creative (see Haaf & Rouder, 2017, 2018 for a discussion of individual differences).

In order to examine the evaluation phase of the AUT, we focus on the decision-making process of whether an alternative use is creative or not. Previous work on creative idea evaluation has primarily focused on how mental or emotional states, or instructions lead to better judgements of how creative ideas are, whether of people's own ideas or those of others (de Buisonjé, Ritter, de Bruin, ter Horst, & Meeldijk, 2017; Grohman et al., 2006; Herman & Reiter-Palmon, 2011; Mastria, Agnoli, & Corazza, 2019; Puente-Diaz, Cavazos-Arroyo, & Puerta-Sierra, 2021; Rietzschel, Nijstad, & Stroebe, 2010; Ritter et al., 2012; Runco & Smith, 1992; Silvia, 2008). In contrast, we use a process modeling approach to better understand the cognitive underpinnings of creativity evaluation and apply the commonly used two-choice response time (RT) paradigm from the decision making literature (e.g., Krypotos, Beckers, Kindt, & Wagenmakers, 2015; Ratcliff, 1978; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), here in the form of the Creative-or-Not task (CON-task). For example, given the object "book" one must decide, as fast as possible, whether using it as a "roof tile" is creative (or not). The CON-task stimuli vary on the two dimensions of originality and utility, allowing us to unobtrusively assess individuals' implicit values of originality and utility when judging creativity in the CON-task. Our task is somewhat similar to the creativity evaluation test (Benedek et al., 2016), a test to detect individual differences in creative evaluation ability amongst prospective secondary school teachers, as we also present others' ideas on the AUT to participants. However, where their focus was on assessing evaluation ability, we aim

to explore the decision-making process and individual differences in implicit conceptions of creativity. As such, responses are not correct or incorrect, but just the respondent's opinion.

To gain insight into the cognitive basis of the evaluation process, we model CON-task data using the drift diffusion model (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008; Wabersich & Vandekerckhove, 2014). The DDM is a cognitive model of the processes during two-choice decisions (Ratcliff & McKoon, 2008). The model essentially assumes that people make decisions through noisy evidence accumulation (see Figure 1). Over time, the individual supposedly accrues more and more information about a stimulus to answer a posed question and then ultimately crosses a decision boundary, for example to decide that "Yes, roof tile is a creative use for a book". For every new sample of information, the individual determines whether it is in line with the "Yes" or the "No" decision and thereby sequentially adds new to the old evidence (Wagenmakers, 2009). This information accumulation process ends when a certain threshold of evidence is reached.

The DDM has commonly been applied to studies of language, perception, or memory retrieval (e.g., Ratcliff, 1978, 2002; Ratcliff, Gomez, & McKoon, 2004) where participants make simple and fast decisions that are either correct or incorrect (e.g., Is CFREE a word or not? Meyer & Schvaneveldt, 1971). But it has also been used to model longer, value-based choices where there is no objectively correct response (Hutcherson, Bushong, & Rangel, 2015; Krajbich, Armel, & Rangel, 2010; Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010). For example, Milosavljevic et al. (2010) showed that the DDM can computationally describe decisions in a binary food choice task. For two-choice decisions where a simple random walk is too simplistic, the DDM might not be complex enough. However, numerous complex decision-making models can be reduced to the DDM (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Since the cognitive processes underlying creativity evaluation are largely unknown, we believe that the DDM serves as a good starting model for the evaluation phase of the AUT – also due to its useful psychometric properties: the DDM can be linked to the two parameter logistic (2PL)

model, a classic item response theory model to measure individual differences (Tuerlinckx & Boeck, 2005; van der Maas et al., 2011). Furthermore, it separates potentially confounding processes related to stimulus encoding or motor control.

Taking into account both responses and RTs, the DDM allows us to explore the role of speed in creativity evaluation as assessed by the CON-task. The originality and utility of the CON-task stimuli may not only influence decision outcomes, but also decision speed. RTs tend to increase with difficulty (Ratcliff & McKoon, 2008). Accordingly, the more clearly creative a CON-task stimulus may be perceived, the faster it would be judged. Since highly original and useful ideas tend to be considered creative, they might be more easily evaluated regarding creativity than medium original ones and they might be judged faster. Using this logic, highly unoriginal stimuli might also be evaluated faster than medium original ones. Altogether, this would suggest an inverted u- or even v-shaped relationship between originality and RT. The same might be the case for utility. In this paper, we explore this idea by measuring both decision outcomes and decision speed in the CON-task.

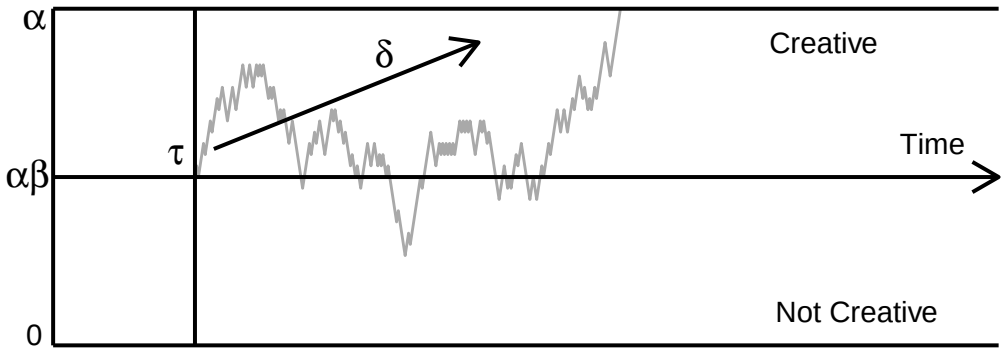

Figure 1. A graphical illustration of the DDM; α = boundary separation, indicating the evidence required to respond 'Creative' or 'Not creative'; β = initial bias to choose one response over the other; δ = average drift rate, indicating the rate of information accumulation; τ = non-decision time, indicating the time used for processes apart from the actual decision-making.

As a plausibility check for our cognitive modeling approach, we also examine how well

individuals' implicit values and their explicit opinions about the relevance of the two dimensions for creativity are aligned. Furthermore, we investigate whether those who value originality (or utility) also tend to produce more original (or useful) AUT responses, respectively. Regarding the latter, we build on Caroff and Besançon (2008)'s study who found a positive relationship for originality and called for research examining the same for "appropriateness" (i.e., a term often used as synonym for the utility dimension).

To this end, we conducted two studies. Study 1 is exploratory where we initially fit the DDM model. In Study 2 we validate our pre-registered model from Study 1 and test specific hypotheses based on Study 1's results.

General Method

Bayesian Hierarchical Diffusion Modeling of the CON-task

The DDM conceptualizes the response process in the CON-task as an interaction of several unobservable cognitive processes (Ratcliff, 1978; Ratcliff & McKoon, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011). Each of these is represented by a parameter (see Figure 1). We use the simplest complete version of the DDM comprising four parameters (Ratcliff, 1978; Ratcliff & McKoon, 2008; Wabersich & Vandekerckhove, 2014). First, the model assumes that the decision whether a use is creative or not is initially determined by β, which reflects the a priori bias towards either choice, regardless of stimulus characteristics. Applied to the CON-task, this is an initial preference for "Yes, creative" or "No, not creative". Second, according to the DDM, individuals gradually extract and accumulate noisy information from the stimulus regarding its creativity, which in turn determines the drift rate δ, the tendency to respond "creative" or "not creative". Positive values suggest a drift towards the upper boundary and negative ones a drift towards the lower boundary. Drift rates around zero suggest that a stimulus is perceived as ambiguous. The higher the absolute drift rate, the easier and faster the creativity evaluation, and the stronger the

evidence for the decision. The evidence accumulation ends when either of the two decision boundaries is reached. Third, the boundary separation parameter α reflects the distance between the two response boundaries and can be interpreted as response caution, where more hesitant creativity judges have a greater boundary separation. Finally, the parameter τ refers to the non-decision time. This parameter captures the processes taking place before and after the actual decision process such as stimulus encoding and motor control processes.

In this paper, the most central DDM parameter is the drift rate. The assumption is that stimulus originality and utility both positively affect the drift rate in that the more original and useful a CON-task stimulus is, the greater the tendency to respond "creative". Moreover, the drift rate is the only model parameter that is influenced by stimulus characteristics because the remaining parameters are already set before the creativity evaluation starts (e.g., Vandekerckhove et al., 2011).

Bayesian hierarchical modeling. We estimated the model in a Bayesian hierarchical framework (Lee, 2011; Rouder & Lu, 2005; Vandekerckhove et al., 2011), allowing us to examine the data both at the population-level and at the individual-level. Hierarchical modeling provides rather conservative estimates of individual differences because it shrinks the individual effects towards the population mean (e.g., Efron & Morris, 1977; Haaf & Rouder, 2018).

We chose Bayesian estimation for three reasons. First, even without a hierarchical extension, applying the DDM to data is computationally expensive (e.g., Tuerlinckx, 2004). Extending it hierarchically makes the model quickly intractable when using the frequentist approach of maximum likelihood estimation (Vandekerckhove et al., 2011). Second, Bayesian inference has several advantages such as an intuitive treatment of uncertainty regarding the model parameters (Wagenmakers, 2009). Third, Bayesian hierarchical modeling is the preferred method for small trial numbers as simulation studies suggest that this method can recover individual variation relatively successfully even with small numbers of observations per participant (Ratcliff & Childers, 2015).

Model Specification

A detailed and complete model specification of the DDM used in Study 1 and 2 can be found in Appendix A. Here, we describe how we decomposed the drift rate parameter and the hierarchical structure of the model. To explore the influence of originality and utility when judging creativity in the CON-task, we regressed the drift rate on the originality and utility ratings of the stimuli. In both studies, we included random intercepts and random slopes to explore individual differences. Furthermore, because the response times and proportions of "creative" responses vary considerably across the 64 CON-task stimuli, we also included random intercepts pertaining to the stimuli.

In both studies, we decomposed the drift rate as follows. Let $\delta_{(ij)}$ denote the drift rate for the ith participant, $i = 1, ..., I$, in the jth trial or stimulus, $j = 1, ..., 64$, of the CON-task, then

$$\delta_{(ij)} = \theta_{\delta(i)} + \phi_{\delta(j)} + \theta_{OR(i)} z_{OR(j)} + \theta_{UT(i)} z_{UT(j)}.$$

The parameters $\theta_{\delta(i)}$, $\theta_{OR(i)}$, $\theta_{UT(i)}$, and $\phi_{\delta(j)}$ reflect the drift rate decomposition. Specifically, $\theta_{\delta(i)}$ denotes the drift rate intercept, representing individual i's drift rate for stimuli with average originality and utility ratings. $\phi_{\delta(j)}$ is stimulus j's deviation from the drift rate intercept. Furthermore, $z_{OR(j)}$ and $z_{UT(j)}$ are z-scores of the originality and utility of stimulus j. Lastly, $\theta_{OR(i)}$ denotes the originality effect, and $\theta_{UT(i)}$ the utility effect of individual i on the drift rate.

For most of the remaining DDM parameters, we also incorporated random effects to examine individual differences. In particular, we allowed the boundary separation and the bias parameter to vary across individuals. However, because in Study 1 we encountered identifiability issues when estimating random effects for β, we fixed the bias at the population level. Another exception is that we constrained the non-decision parameter, τ, to

be constant across participants in both studies because interpreting random effects for this parameter has shown to be problematic (Singmann, 2018b).

To examine the interplay of the DDM parameters across participants, we also allowed the random effects pertaining to individuals to be correlated. As such, we assume that the individual effects are drawn from the same multivariate normal distribution with population means $[\mu_\delta, \mu_{\theta_{OR}}, \mu_{\theta_{UT}}, \mu_\alpha, \mu_\beta]^T$ and a variance-covariance matrix $\mathbf{\Sigma}$, i.e.,

$$\begin{bmatrix} \theta_{\delta(i)} \\ \theta_{OR(i)} \\ \theta_{UT(i)} \\ \alpha_{(i)} \\ \beta_{(i)} \end{bmatrix} \sim \text{Multivariate-Normal} \left(\begin{bmatrix} \mu_\delta \\ \mu_{\theta_{OR}} \\ \mu_{\theta_{UT}} \\ \mu_\alpha \\ \mu_\beta \end{bmatrix}, \mathbf{\Sigma} \right).$$

$\mathbf{\Sigma}$ allows for correlations across the random effects pertaining to the individuals. The random effects of the stimuli are orthogonal to the individual random effects. They are also assumed to be randomly sampled from a population distribution (of stimuli),

$$\phi_{\delta(j)} \sim \text{Normal}(0, \sigma_{\delta_\phi}),$$

where 0 is the mean and σ_{δ_ϕ} is the standard deviation.

Since we estimated the model in the Bayesian framework, we needed to specify a prior distribution for each parameter. For Study 1, we used weakly informative priors that restricted the parameter space to a plausible range (see Appendix A). For Study 2, we used the insights gained from Study 1 and specified informative priors to test hypotheses in the Bayesian setting. We discuss these prior choices in the Methods, with details in Appendix A.

Study 1

Data Collection Procedure and Materials

The data were gathered as part of a joint data collection effort by the Department of Psychology at the University of Amsterdam that took place over six sessions within one month. The sessions were filled with questionnaires and tasks from different researchers. Participation in each session was optional and participants received course credit. The full sample in our study consisted of 299 first year psychology students. The age range was 17-41 years ($M = 20.38$, $SD = 2.59$). Participants completed the tasks (i.e., the Alternative Uses Task and the CON-task) and questions all in one session in the order listed below[1].

Alternative Uses Task. Participants completed a computerized version of the Alternative Uses Task (AUT; Guilford, 1967) used to assess their divergent thinking performance. The name of an object was presented on the screen, and participants had two minutes to type as many creative uses for the object as possible (e.g., the use "bath toy" for the object "brick"). During the session, participants were asked to generate uses for two objects, either "brick" and "fork", "fork" and "paperclip". or "paperclip" and "towel". The pairs were counter-balanced over participants. Generated solutions were listed on the screen and new ones were continuously added. Two independent raters who were unaware of the research questions/hypotheses of this study separately scored participants' answers with respect to originality and utility on a five-point scale (1 = not original/useful, 5 = very original/useful). Invalid responses were coded as 0. To assess interrater reliability, we computed the intraclass correlation coefficient (ICC)[2]. The ICC for the originality scores

[1] The Alternative Uses Task was administered in at least one other study during the six testing sessions but no other study made use of the CON-task.

[2] Since we considered both responses and raters as random effects and considered consistency in ratings more important than absolute agreement, we used a twoway model and computed single score ICCs of the type consistency using the irr package throughout this paper (Gamer, Lemon, & Singh, 2019). Note that to

were 0.65 95%CI [0.63, 0.67], 0.68 95%CI [0.67, 0.70], 0.73 95%CI [0.71, 0.76] and 0.67 95%CI [0.63, 0.70] for "brick", "fork", "paperclip" and "towel" respectively. For the utility scores, the corresponding ICCs were 0.56 95%CI [0.54, 0.58], 0.61 95%CI [0.59, 0.63], 0.83 95%CI [0.82, 0.85] and 0.68 95%CI [0.65, 0.71] for "brick", "fork", "paperclip" and "towel" respectively. As performance indicators, we used the mean originality and mean utility score across raters, objects, and responses.

Creative-or-not Task. Participants completed 64 trials of the CON-task. The instructions were in Dutch and read the following: "In a moment, you'll see other people's answers to the 'Creative Uses task'. We would like to know if you think the answers are creative or not creative. Decide as quickly as you can. We will do this task four times, each time with a different object (such as book). You will be shown 16 ideas for each object." On each trial, they were asked "Do you think this use for [object] is creative?", followed by a specific use. Importantly, participants were not instructed regarding the criteria they should apply when deciding whether they find a use creative or not. RTs as well as responses ("creative" or "not creative") were recorded. Trials automatically counted as missing when participants did not answer within nine seconds. The stimuli used had been selected from a collection of AUT responses. Their originality and utility had been independently scored on a scale from 1 to 5 by two creativity researchers. The ICC was 0.88 95%CI [0.80, 0.92] for originality and 0.65 [.49, 0.77] for utility. As stimulus ratings, we took the average originality and utility rating, respectively, across raters. The mean originality rating of the stimuli was $M = 2.98$ ($SD = 1.20$), and the mean utility rating was $M = 3.37$ ($SD = 1.07$). Stimulus originality and utility were negatively correlated, r = -0.61, 95% credible interval (CrI) [-0.91, -0.45], $BF_{10} = 1.43 \times 10^{63}$. The correlation is representative of the trade-off between

compute the ICCs for the AUT objects separately, we used all collected AUT responses during the testing sessions and not only the responses of participants who also completed the CON-task.

[3] Here and for all subsequently reported correlations, we conducted Bayesian correlation analyses using the Bayes factor package including the default prior scale (Morey & Rouder, 2018). Specifically, we used Bayes

the two dimensions "originality" and "utility" that is often found in AUT responses (e.g., Rietzschel et al., 2010; Runco & Charles, 1993).

Importance Ratings of Originality and Utility. After the CON-task, participants indicated, separately, to what extent they thought utility, innovativeness, originality and appropriateness played a role when evaluating creativity (1 = not important at all to 5 = very important).

Results Study 1

The data cleaning is described in Appendix B. The cleaned dataset comprised 293 participants and 17806 trials. The mean RT across participants and trials after excluding data was 1.86 seconds ($Median = 1.61$, $SD = 0.92$). The overall RT distribution is shown in Appendix C. The overall percentage of "creative" responses was 53.83%. The mean RT for "creative" responses across participants and trials was 1.87 s ($SD = 0.93$), and 1.86 s ($SD = 0.92$) for "not creative" responses. Figure 2A shows that RTs for some stimuli were longer than for others. A Bayesian correlation analysis with median-split stimulus data suggested weak evidence for a correlation between RT and originality in the high-utility stimulus group, $r = 0.33$, 95%CrI [0.04, 0.57], $BF_{10} = 4.70$, and anecdotal evidence for no relationship between originality and RT in the low-utility stimulus group[4], $r = 0.17$, 95%CrI [-0.19, 0.50],

factors to quantify the evidence for a correlation ($H_1 : \rho \neq 0$) as opposed to no correlation ($H_0 : \rho = 0$;) and report it together with the posterior mean of the correlation coefficient and the corresponding credible interval.

[4] The exact results depended on whether the median of the stimuli's utility ratings was included in the low- or the high-utility group. When the median was assigned to the low-utility group, the evidence for the correlation between originality and RT in the high-utility group was even smaller: $r = 0.31$, 95%CrI [-0.03, 0.59], $BF_{10} = 2.40$. The evidence for no correlation between originality and RT in the low-utility group was also even smaller to the point where there was practically neither evidence for the presence nor for the absence of a correlation, $r = 0.21$, 95%CrI [-0.10, 0.49], $BF_{01} = 1.06$

$BF_{01} = 1.47$ (see Figure 2A). Furthermore, Figure 2B and a Bayesian paired t-test analysis suggested that participants, on average, responded equally fast with "not creative", $M = 1.88$ s, $SD = 0.39$ s, as opposed to "creative" $M = 1.89$ s, $SD = 0.43$ s, $BF_{01} = 11.32$.

Model Fit. We fitted the DDM using the R package *brms* (Bürkner, 2018) which works with Stan to draw samples from the posterior distribution of Bayesian models (Carpenter et al., 2017). We ran 4 chains with 5000 iterations each. 1500 iterations per chain were used as warmup to adapt the sampler. Consequently, our analyses were based on a total of 14000 iterations.[5]

We decided to not allow the bias parameter β to vary across individuals because when we did, the random effects of the bias and drift rate intercept were highly correlated, suggesting identifiability issues. We therefore estimated the DDM with only a population-level bias parameter. The remaining DDM parameters were not affected by this change.

We performed several model diagnostics procedures and inspected the model fit. There were no signs of non-convergence, with 0 divergent transitions and $\hat{R}$ values (Gelman &

[5] For all analyses, we used R (Version 3.6.1; R Core Team, 2019) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *bayesplot* (Version 1.7.0; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.10.0; Bürkner, 2017, 2018), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *corrplot2017* (Wei & Simko, 2017), *cowplot* (Version 1.0.0; Wilke, 2019), *DescTools* (Version 0.99.29; al., 2019), *dplyr* (Version 1.0.0; Wickham, François, Henry, & Müller, 2020), *ggplot2* (Version 3.3.5; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *here* (Version 0.1; Müller, 2017), *irr* (Version 0.84.1; Gamer, Lemon, & <puspendra.pusp22@gmail.com>, 2019), *lpSolve* (Version 5.6.15; Berkelaar & others, 2020), *magrittr* (Version 2.0.1; Bache & Wickham, 2020), *Matrix* (Version 1.2.17; Bates & Maechler, 2019), *msm* (Version 1.6.9; Jackson, 2011), *mvtnorm* (Version 1.0.11; Genz & Bretz, 2009), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *psych* (Version 1.8.12; Revelle, 2018), *Rcpp* (Version 1.0.7; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2018), *rstan* (Version 2.19.2; Stan Development Team, 2019a), *StanHeaders* (Version 2.19.0; Stan Development Team, 2019b), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.1.5; Müller & Wickham, 2021), and *tidyr* (Version 1.0.0; Wickham & Henry, 2019).

Rubin, 1992) below 1.01 (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020). Additionally, we assessed the model fit using posterior predictive checks (see the online supplementary material). Overall, apart from some misfit in the outer quantiles of the RT distribution, the DDM could reproduce the data quite accurately and appeared to provide an acceptable account of the data.

Modeling Results. A summary of the estimated fixed and random effects parameters can be found in Table 1 and 2, respectively, and a summary table with the random effects correlation parameters in Appendix B.

Regarding the fixed effects, our main focus of interest was on $\mu_{\theta_{OR}}$ and $\mu_{\theta_{UT}}$, the overall effects of stimulus originality and utility on the drift rate δ. For $\mu_{\theta_{OR}}$, the posterior mean was 0.41, and the 95% CrI was $[0.31, 0.50]$. For $\mu_{\theta_{UT}}$, the posterior mean was 0.10, and the 95% CrI was $[0.00, 0.19]$. Both posterior means were positive. However, while the 95% credible interval (CrI) of $\mu_{\theta_{OR}}$ did not include zero, the CrI of $\mu_{\theta_{UT}}$ was very close to zero.

In general, all estimated posterior means of the remaining fixed effects parameters seem plausible as the CrIs were rather narrow and the parameters lie within a reasonable range (see Table 1). On average, there was no a-priori bias towards the response options "creative" or "not creative" (see μ_β in Table 1). This suggests that participants were on average equally likely to choose either of the two response options before stimulus onset. However, the boundary separation, or response caution, was higher than found in most applications of the diffusion model (e.g., Matzke & Wagenmakers, 2009, see μ_α in Table 1). One explanation for this rather high value is that the RTs in the CON-task were considerably slower than RTs in tasks typically modeled by the diffusion model.

We found substantial individual differences in all variability parameters. Notably, the results showed substantial variability across participants in the originality and utility effects on the drift rate. The posterior means of σ_{OR} and σ_{UT} were 0.23 and 0.19, respectively. The posterior means and credible intervals of all variability parameters are listed in Table 2. None of the CrIs included zero suggesting considerable variability across stimuli and across

participants.

Figures 3A and B visualize this variability in the originality and utility effects on the drift rate by depicting the posterior means of the individual originality slopes $\theta_{OR(i)}$ and utility slopes $\theta_{UT(i)}$ and their corresponding CrIs in increasing order. The figures show substantial individual differences. Regarding the individual originality effects, there were even 11 participants with a negative posterior mean (3.75%). However, the 95%CrI of these estimates included zero. In total, the CrI of 54 individuals included zero (18.43%), suggesting that, at the very least, for some individuals, the effect of stimulus originality on the drift rate was weaker and for some it was stronger.

Participants further differed in their utility slopes. As shown in Figure 3B, for some, the utility effect on the drift was around zero, for some it was positive, and for a few the effect was even negative. Specifically, the majority of the individual CrIs (n = 207; 70.65%) included zero, 80 (27.30%) of them excluded and were above zero, and 6 (2.05%) individual CrIs excluded and were below zero.

Individual differences also manifested themselves in a negative correlation between the originality and utility slopes. Here the posterior mean of $\rho_{\sigma_{\theta_{OR}}\sigma_{\theta_{UT}}}$ was -0.44, 95%CrI [-0.60, -0.27], and the correlation between the individual originality and utility slopes, based on the posterior means, was r = -0.67, 95%CrI [-0.73, -0.61], $\text{BF}_{10} = 6.33 \times 10^{37}$. It is expected that r is greater than ρ because $\rho_{\sigma_{\theta_{OR}}\sigma_{\theta_{UT}}}$ is a population parameter taking uncertainty into account and r reflects the data in our sample. The greater the individual effect of stimulus originality on the drift rate, the smaller the effect of stimulus utility and vice versa. This correlation could also be explained by the substantial negative correlation between the originality and utility ratings for the CON-task stimuli (r = -0.61). We ruled this out by re-estimating the DDM when excluding the stimuli that contributed the strongest to the negative correlation. Specifically, we excluded the data from 20 items leaving us with CON-task data based on 44 stimuli. Excluding those items reduced the stimulus originality-utility ratings correlation from r = -0.61 to r = -.15. Despite this reduced

correlation across stimuli, the negative correlation across originality and utility effects on individual drift rates remained substantial, r = -.49, 95%CrI [-0.67, -0.29] (see supplementary materials). This result suggests that the correlation across effects is not (solely) a function of stimulus characteristics.

Figure 4A, shows the multivariate, joint posterior distribution of the originality and utility effects and Figure 4B the individual posterior means and corresponding standard deviations to visualize this correlation. An overview table of the correlations among all random effects parameters can be found in Table A1 in Appendix D. Figure 4B further depicts two clusters that are the results from an exploratory k-means cluster analysis: it seems that one cluster comprises individuals with a positive effect of utility and a rather small effect of stimulus originality, and the other one individuals with a stimulus utility effect around zero and a positive effect of originality.

CON-task and self-report ratings of originality and utility. As a plausibility check for our rationale behind the drift rate regression, we examined whether the self-reported importance ratings of originality, innovativeness, utility, and appropriateness corresponded to the originality and utility effects on the drift rate. We summed up the ratings of appropriateness and utility and innovativeness and originality, respectively. There was a positive correlation between participants' importance ratings of originality and the posterior means of their originality slopes, $r = 0.32$, 95%CrI [0.21, 0.42], $BF_{10} = 1.04 \times 10^6$ and between their ratings of utility and utility slopes, $r = 0.36$, 95%CrI [0.26, 0.45] , $BF_{10} = 1.66 \times 10^8$. The more participants indicated that originality was important when determining whether something is creative or not, the greater their influence of stimulus originality on their drift rate. The more they indicated that utility was important, the greater their effect of stimulus utility on their drift rate. There were also negative correlations between the originality importance ratings and the utility slopes, r = -0.33, 95%CrI [-0.43, -0.23] , $BF_{10} = 4.39 \times 10^6$, and between the utility ratings and the originality slopes, r = -0.36, 95%CrI [-0.46, -0.26], $BF_{10} = 1.90 \times 10^8$.

CON-task and divergent thinking. Given the substantial variability in the stimulus originality and utility effects on the drift rate, we explored whether this variability was related to variability in divergent thinking performance as assessed by the AUT. To this end, we computed correlations among the individual posterior means and participants' AUT performance scores. The data cleaning for the AUT task is described in Appendix B.

We found a positive correlation between AUT originality scores and the originality slope posterior means, $r = 0.18$, 95%CrI [0.06, 0.30] , $BF_{10} = 8.41$, suggesting that the more original the AUT responses, the greater the influence of originality on the drift rate in the CON task. We also found a negative correlation between the posterior means of the utility slopes and participants originality scores, r = -0.16, 95%CrI [-0.28, -0.04], $BF_{10} = 3.11$. The more original responses participants produced in the AUT, the smaller their effects of stimulus utility on the drift rate. However, there was no correlation between the AUT utility scores and the stimulus utility effects on the drift rate, $r = 0.02$, 95%CrI [-0.11, 0.15] , $BF_{01} = 6.15$, and also no correlation between AUT utility scores and stimulus originality effects, r = -0.03, 95%CrI [-0.16,0.10], $BF_{01} = 5.82$.

Since the application of the DDM to creativity is novel and since Study 1 was conducted in a rather exploratory manner, we aimed to assess the robustness of our findings in a second, preregistered Study (https://osf.io/7gt45/). In Study 2, we therefore specified hypotheses based on Study 1's results as well as previous research and re-fitted the DDM on an independent dataset.

Study 2

Based on what we had learned from Study 1 and based on previous research, we expected positive effects of stimulus originality (H1) and stimulus utility (H2) on the drift rate. We also expected the effect of stimulus originality to be larger than the effect of stimulus utility (H3). Furthermore, given the observed substantial negative correlation between the individual stimulus originality and utility effects in Study 1, we expected a

negative correlation among those effects in Study 2 (H4). Specifically, we hypothesized that the greater an individual's effect of stimulus originality on the drift rate, the smaller the effect of stimulus utility would be and vice versa. Since in Study 1, we observed substantial individual differences in the extent to which stimulus originality and utility influenced the drift rate, we also expected non-zero variability across individuals in those effects (H5a and H6a).

Previous research suggests that originality plays a superior role in creativity judgments compared to utility (e.g., Caroff & Besançon, 2008; Diedrich et al., 2015; Runco & Charles, 1993). Based on these findings, we tested the hypothesis that everyone would have a positive effect of stimulus originality on the drift rate (H5b). Note that in Study 1, a few participants seemed to have a negative effect of originality. However, because this finding concerned only few participants and because we were not aware of any theory or research supporting it, we did not consider it robust enough to inform Study 2. Instead we decided to quantify the evidence for the ordinal constraint that everyone has a positive effect in Study 2 (Haaf & Rouder, 2017, 2018). Regarding the effects of stimulus utility on the drift rate, we expected that some individuals would have a positive effect, some a negative effect, and some no effect (H6b). Given Study 1's results, we also expected that individuals' stimulus originality and utility effects would be positively associated with their self-reported importance ratings of originality and utility for creativity. More specifically, we expected individual originality effects on the drift rate would increase as the self-reported importance ratings of originality increase (H7a), and individual utility effects on the drift rate would increase as the importance ratings of utility increase (H7b).

Finally, we specified hypotheses regarding the association between CON-task judgements and AUT performance. We expected that originality scores on the AUT would be positively correlated with stimulus originality effects and negatively correlated with stimulus utility effects on drift rates (H8a and H8b). Similarly, we expected that AUT utility scores would be positively correlated with stimulus utility effects on CON-task drift rates

and negatively correlated with stimulus originality effects (H9a and H9b). Note that Study 1 did not support H9a and H9b. However, in both cases, we did not consider the evidence for the null to be convincing. Based on common sense, we still expected that the more useful one's AUT responses are, the more one values utility when judging creativity and the more one disregards originality – also since studies suggest that the more creative someone is, the better they are at judging creativity (e.g., Benedek et al., 2016; Silvia, 2008). Study 2 served as a robustness check for this belief.

Data Collection Procedure and Materials

As in Study 1, data collection was centrally organized by the faculty of Psychology at the University of Amsterdam and took again place over different sessions. All tasks related to creativity were administered during the same session in the order listed below. In total, 172 first-year psychology students completed the CON-task, the age range was 17-47 years, ($M = 20.50$, $SD = 2.98$, and participants again received course credit for their participation.

Alternative-Uses-Task Participants completed the AUT (Guilford, 1967) for the objects "brick" and "paperclip" and were again given two minutes for each object. Two independent raters who were unaware of the research questions/hypotheses of this study again separately scored participants' answers with respect to originality and utility on a five-point scale and coded invalid responses as zero. Interrater reliability as assessed by ICCs[6] can be considered moderate to good: for the object "brick", the ICC was 0.67 95%CI [0.65, 0.69] for originality and 0.66 95%CI [0.63, 0.68] for utility. For the object 'paperclip the ICC was 0.77 95%CI [0.75, 0.79] for originality and 0.71 95%CI [0.68, 0.73] for utility. As performance indicators, we again used the mean originality and mean utility score across

[6] As in Study 1, we used a twoway model and computed single score ICCs of the type consistency using the irr package (Gamer, Lemon, & Singh, 2019). Note that to compute the ICCs for the objects separately, we again used all collected AUT responses and not only the responses of the participants who also completed the CON task.

raters, objects, and responses.

Creative-Or-Not task Participants again completed the same 64 stimuli as the participants in Study 1.

Self-reported importance-ratings On four separate items, participants again indicated after the CON-task how important they thought originality, innovativeness, appropriateness, and utility were when deciding whether something is creative or not.

Results Study 2

The hypotheses and analysis plan for Study 2 were preregistered before seeing the data. We employed the same exclusion criteria as in Study 1 (see Appendix B).

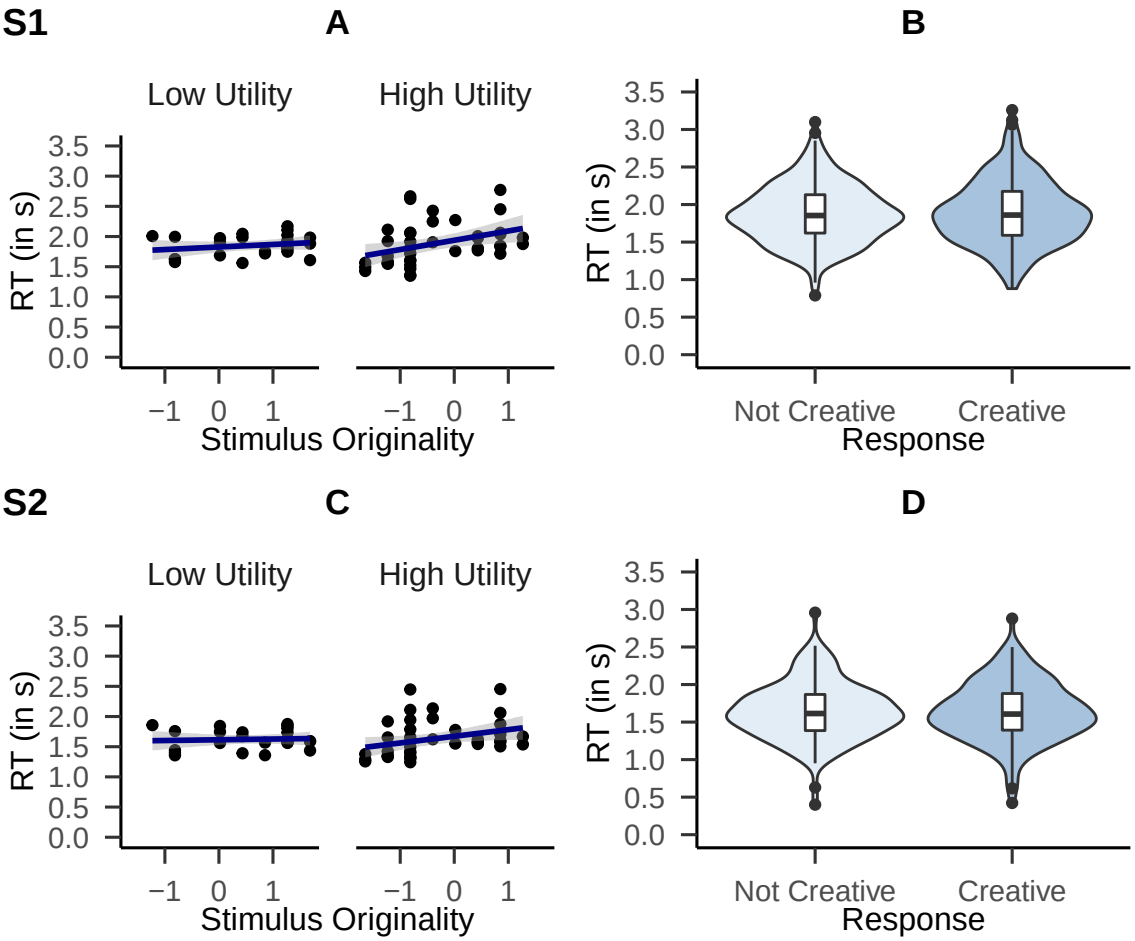

Figure 2. A and C show the mean stimulus response times as a function of stimulus originality and utility in Study 1 and 2, respectively. Each dot represents a stimulus. Low and high utility stimuli were categorized by median split. B and D show the distribution of mean response times for participants' 'creative' and 'not creative' decisions.

The descriptives were very similar to Study 1. The mean RT across participants and trials was 1.62 s and the median 1.44 s ($SD = 0.79$). The overall RT distribution is shown in Appendix C. The overall proportion of "creative" responses was 0.57. The mean RT for "creative" responses was 1.62 s ($SD = 0.78$). For "not creative" responses, this was 1.63 s ($SD = 0.79$). Stimuli with higher scores on originality and utility were again answered more slowly (see Figures 2C). A Bayesian correlation analysis with median-split data suggested weak evidence for a correlation between RT and originality in the high-utility group, $r =$

0.27, 95%CrI [-0.01, 0.53], $BF_{10} = 2.09$, and also weak evidence for no association between originality and RT in the low-utility group[7], $r = 0.05$, 95%CrI [-0.31, 0.40], $BF_{01} = 2.18$. Furthermore, as in Study 1, Figure 2D and a Bayesian paired t-test analysis suggested that participants, on average, responded equally fast when they answered "not creative", (M = 1.65 s, $SD = 0.38$ s), as opposed to "creative" $M = 1.64$ s, $SD = 0.40$ s, $BF_{01} = 10.96$.

Model Fit. In contrast to Study 1, we used informative priors for the stimulus originality and utility effects on the drift rate, based on Study 1's estimation results. Specifically, we specified truncated normal distributions as priors for $\mu_{\theta_{OR}}$, and $\mu_{\theta_{UT}}$,

$$\mu_{\theta_{OR}}, \mu_{\theta_{UT}} \sim \text{Normal}^{+}(0, 0.2),$$

.

All remaining priors were the same as in Study 1 except for β (see Appendix A). We again fitted the model using the R package *brms* (Bürkner, 2018), ran 4 chains with 4000 iterations of which 500 iterations per chain were used as warmup, leaving us with 14000 iterations to base the analysis on.

We again inspected the model diagnostics and model fit. There were no signs of non-convergence, with 0 divergent transitions, and $\hat{R}$ values below 1.01 (Vehtari et al., 2020). Moreover, we again assessed the model fit using posterior predictive checks. The model fit was similar as in Study 1 and overall acceptable (see the online supplementary material). As a robustness check, we also re-estimated the model including participants that we excluded based on exclusion criterion 3 (fewer than 47 remaining trials; see the online supplementary materials).

---

[7] The exact results again depended on whether the median of the stimuli's utility ratings was included in the low- or the high-utility group. When the median was assigned to the low-utility group, the evidence for the correlation between originality and RT in the high-utility group was even smaller: $r = 0.27$, 95%CrI [-0.07, 0.56], $BF_{10} = 1.52$. However, the Bayes factor in favor of no correlation between originality and RT in the low-utility group was slightly bigger, $r = 0.10$, 95%CrI [-0.21, 0.39], $BF_{01} = 2.21$

Modeling Results. Summary statistics of the estimated model parameters are shown in Table 1 and 2. A table with the correlations among the random effects can be found in Appendix B. Overall, the estimated DDM parameters were very similar to the ones in Study 1.

Table 1

Posterior mean, standard deviation of the posterior distribution, 95% credible interval and $\hat{R}$ statistic for the fixed effects (population-level) parameters

	Study 1				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
μ_δ	0.06	0.04	-0.01	0.14	0.15	0.05	0.05	0.26
μ_β	0.49	0.00	0.49	0.50	0.49	0.01	0.48	0.50
$\mu_{\theta_{OR}}$	0.41	0.05	0.31	0.50	0.40	0.05	0.30	0.51
$\mu_{\theta_{UT}}$	0.10	0.05	0.00	0.19	0.10	0.05	0.01	0.21
μ_α	2.96	0.03	2.91	3.01	2.72	0.04	2.64	2.80
τ	0.28	0.00	0.27	0.28	0.29	0.00	0.29	0.29

Note. $\mu_{\theta_{OR}}$, $\mu_{\theta_{UT}}$, and μ_δ are standardized estimates as the originality and utility ratings of the stimuli are z-scores. SD = standard deviation; LB = lower bound; UB = upper bound.

Table 2

Posterior mean, standard deviation of the posterior distribution, 95% credible interval and $\hat{R}$ statistic for the variability parameters. σ_{δ_ϕ} denotes the variability across stimuli

	Study 1				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
σ_{δ_ϕ}	0.27	0.03	0.23	0.33	0.33	0.03	0.27	0.40
σ_{δ_υ}	0.27	0.01	0.25	0.3	0.36	0.03	0.32	0.42
σ_{OR}	0.23	0.01	0.21	0.26	0.20	0.02	0.16	0.24
σ_{UT}	0.19	0.01	0.17	0.22	0.20	0.02	0.17	0.24
σ_α	0.4	0.02	0.36	0.44	0.46	0.03	0.41	0.52
σ_β	-	-	-	-	0.05	0.01	0.04	0.06

Note. SD = standard deviation; LB = lower bound; UB = upper bound.

Hypotheses testing. To test our hypotheses, we used the Savage-Dickey method (Dickey, 1971) of approximating Bayes factors. In this method, the Bayes factor is computed by a ratio of the prior and posterior density at the value zero. Assessing H1, we computed a Bayes factor comparing how well the hypothesis of a positive effect of stimulus originality on the drift rate predicted the data in comparison to the null hypothesis. All posterior samples were greater than zero. Therefore, the evidence in favor of H1 can be regarded as greater than 14000. In contrast, there was mere anecdotal evidence for an effect of utility on the drift rate, the Bayes factor was 1.88. This means that the data was 1.88 times more likely to have occurred under H2 than under the null hypothesis. Not suprisingly, the Bayes factor for H3 that the overall effect of stimulus originality is greater than the effect of stimulus utility

(rather than the other way around), was again greater than 14000.

We examined the correlation between the stimulus originality and utility effects on the drift rate (H4). The random effects correlation as estimated by the model, $\rho_{\sigma_{\theta_{OR}}\sigma_{\theta_{UT}}}$, had a posterior mean of -0.34, 95%CrI $[-0.59, -0.09]$, and the correlation based on the individual posterior means was r = -0.64 95%CrI [-0.73, -0.54], $\text{BF}_{10} = 2.79 \times 10^{16}$. We again explored the robustness of the correlation between the originality and utility effects on the drift rate and by estimating the DDM based on data from a subset of weakly correlated stimuli (see Study 1)[8]. Unlike in Study 1, the correlation between the effects became considerably smaller. The posterior mean of the correlation was reduced to the size of the stimulus correlation in the reduced item pool (r = -.17), suggesting that the negative correlation between originality and utility stimulus effects on the drift rate was not as robust as in Study 1.

The individual differences were again reflected in the variability parameters of the originality and utility effects on the drift rate. The posterior mean of the variability parameters for σ_{OR} was 0.20 and for σ_{UT} this was also 0.20. The Savage-Dickey density ratio cannot be used to compute a Bayes factor quantifying the evidence for individual variability. However, the 95% credible intervals of both variability parameters did not include zero, supporting H5a and H6a that the variability was substantial. To further examine the extent of individual variation, we again plotted individual posterior means and credible intervals in increasing order (see Figures 3C and D). Regarding originality, the CrI of 23 individuals included zero (15.13%) indicating that some individuals' drift rates were only weakly (if at all) determined by the stimuli's originality. Out of these participants, there were even 3 participants with a negative posterior mean (1.97%). Participants also again differed in their utility slopes. The majority of the individual CrIs (n = 118; 77.63%) included zero, 32 (21.05%) of them excluded and were above zero, and 2 (1.32%) individual CrIs excluded and were below zero.

[8] This analysis was not pre-registered.

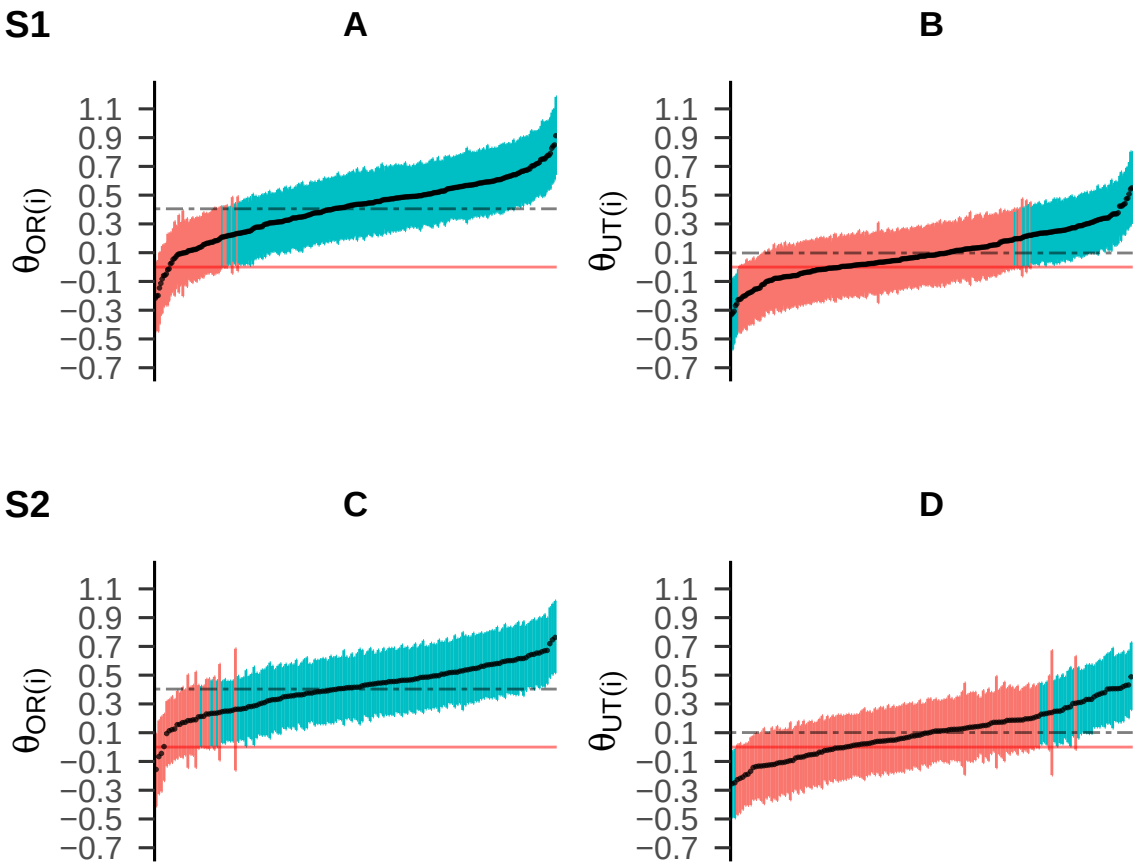

Figure 3. The plots show the posterior means and the 95 % credible interval (CrI) for each participant in increasing order. The dashed horizontal line denotes the population-level posterior means $\mu_{\theta_{OR}}$, $\mu_{\theta_{UT}}$. CrIs colored in red included zero. Plot A and C show the individual estimates of the originality effect in Study 1 and Study 2, respectively. Only a few lower bounds of the CrIs were below zero. Plot B and D show the individual estimates of the utility effect in Study 1 and Study 2, respectively. Some CrIs are above zero, some around zero, and a few below zero.

In line with Haaf and Rouder (2017, 2018), we tested the hypothesis that all individuals have a positive effect of stimulus originality on the drift rate using the encompassing prior approach (H5b; Klugkist, Laudy, & Hoijtink, 2005; Klugkist & Hoijtink, 2007). Here, we compared the predictive accuracy of the hypothesis that everyone's originality slope is positive to the hypothesis that originality effects can be positive, zero or

negative. The constraint that everyone has a positive effect was not fulfilled in any of the iterations. Therefore, the Bayes factor for the latter, unconstrained hypothesis can be considered at least 2,261.16, (assuming that the next iteration might be the first in favor of the ordinal constraint) suggesting that not everyone has a positive originality effect on the drift rate. Regarding the individual utility effects, we tested the hypothesis that some individuals have a positive effect of stimulus utility on the drift rate, some have a negative effect, and some have no effect (H6b). Here, the data provided overwhelming support for the unconstrained hypothesis H6b over the hypothesis that everyone has a positive utility effect as the Bayes factor was again at least 2,261.16 (assuming that the next iteration might be the first in favor of the ordinal constraint).

CON-task and self-report ratings of originality and utility. To test the hypothesis that stimulus originality and utility effects on the drift rate were positively correlated with self-reported importance ratings of originality and utility, respectively (H7a and b), we conducted one-sided Bayesian correlation tests again using sum scores. There was a positive correlation between the originality sum scores and the individual originality effects, $r = 0.17$, 95%CrI [0.03, 0.31], $BF_{10} = 2.56$ as well as a positive correlation between the utility sum scores and the individual utility effects, $r = 0.24$, 95%CrI [0.08, 0.38], $BF_{10} = 24.83$, supporting both H7a and H7b. We again conducted an exploratory cluster analysis on the individual posterior means (see Figure 4), which yielded practically the same result as in Study 1[9].

[9] This analysis was not pre-registered.

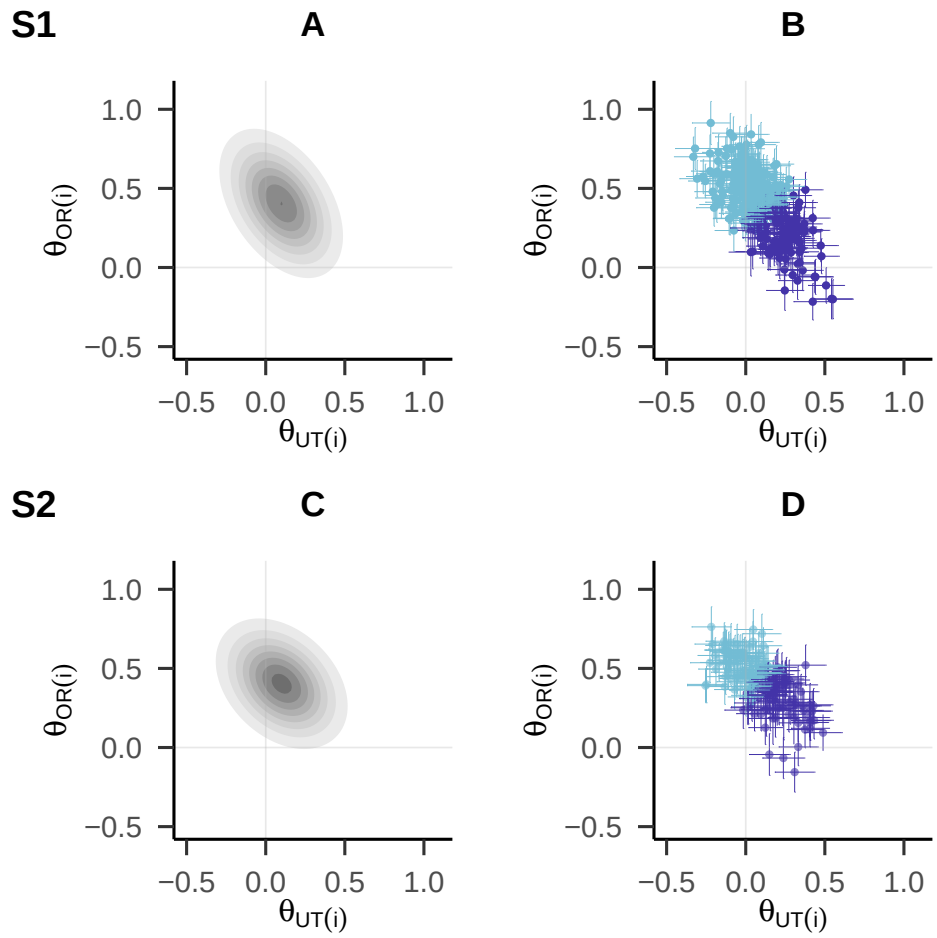

Figure 4. A and C show the multivariate, joint posterior distribution of the originality effects ($\theta_{OR(i)}$) and utility effects ($\theta_{UT(i)}$) on the drift rate, in Study 1 and Study 2, respectively. Darker areas indicate greater density. B and D show the individual posterior means of the originality and utility effects including the standard deviations. Each dot represents a participant. The dots are colored according to the results of k-means cluster analysis. Both plots depict the negative correlation between the originality and utility effects.

CON-task and divergent thinking. To examine associations between the stimulus originality and utility effects on the drift rate and creative performance (H8 and H9), we followed the same data cleaning procedures as described in Study 1. Since we had directed hypotheses, we used one-sided correlation tests. There was a positive correlation between the stimulus originality effects and the AUT originality scores, $r = 0.21$, 95%CrI

[0.05, 0.36], $BF_{10} = 6.32$. However, there was no evidence for a correlation between AUT originality scores and the stimulus utility effects from the CON-task, r = -0.13, 95%CrI [-0.28, -0.01], $BF_{01} = 1.19$ (H8b). Moreover, we considered the evidence to be inconclusive as the data were only 1.19 times more likely to have occured under the null hypothesis rather than hypothesis H8b. Furthermore, there was a positive correlation between the AUT utility scores and the stimulus utility effects, $r = 0.20$, 95%CrI [0.05, 0.36], $BF_{10} = 5.90$ (H9a), and a negative correlation between the AUT utility scores and the stimulus originality effects on the drift rate, r = -0.26, 95%CrI [-0.41, -0.10], $BF_{10} = 34.54$ (H9b).

Discussion

In this paper, we aimed to study the cognitive basis of how people decide whether an idea is creative or not during divergent thinking. Divergent thinking is a two-phase process where people come up with ideas (ideation phase) and decide whether or not to use them (evaluation phase; Basadur, 1995; Guilford, 1967; Runco & Basadur, 1993). We focused on the evaluation phase and used Bayesian cognitive modeling to better understand it. Using a novel application of the drift diffusion model (DDM) we aimed to understand how people weigh the two components of creativity, originality and utility, when deciding whether an idea is creative or not. We used the Creative-or-Not (CON) task, a timed two-choice decision-making task, where people are presented with ideas of how to use an everyday object (e.g., use a book as a roof tile) and then decide whether these are creative or not. The CON-task stimuli varied on how original and how useful they were (e.g., using a book to read is useful, but not original, whereas using a book as roof tile can be both). This way we could estimate individual differences in how originality and utility implicitly contributed to people's creativity judgements by examining their effect on the DDM drift rate parameter. The drift rate represents people's tendency to respond "creative" or "not creative". The higher its absolute value, the greater the accumulated evidence for creative-or-not decisions, and the faster the response. Our main findings were: (1) stimulus originality strongly influenced participants' tendency towards creative responses (i.e., the drift rate), whereas stimulus utility was only somewhat related to it; (2) there were large individual differences in how much participants took the originality and utility of presented ideas into account (i.e., varying individual effects of stimulus originality and utility on the drift rate); (3) participants' implicit tendencies or values of originality and utility on the CON-task were aligned with their self-reported importance ratings of originality and utility for creativity; (4) alternative uses task (AUT) scores for originality and utility coincided with how much the originality and utility of the presented CON-task idea weighed in on their decisions. We now

discuss each of these results in more detail.

Our finding that people value originality when evaluating creativity, evidenced by a substantial overall effect of originality on DDM drift rates, are in line with previous studies showing that people associate originality more so than utility with creativity (e.g., Caroff & Besançon, 2008; Diedrich et al., 2015; Runco & Charles, 1993). In both studies, the utility of ideas was only somewhat related to participants' tendency towards creative responses. In earlier work we also see that utility, or one of its creative synonyms (e.g., appropriateness, value) is less valued when it comes to judging creativity (e.g., Diedrich et al., 2015). However, despite the considerably smaller effect, a Bayesian model comparison supports our proposed model and suggests that utility matters when evaluating the creativity of ideas[10]. Future work with the CON-task could include additional information about presented ideas, such as "surprise" (Boden, 2004) and "impact" (Sternberg & Lubart, 1996), to examine how these components implicitly influence people's creativity judgments.

Interestingly, just as researchers sometimes disagree on the definition of creativity (Simonton, 2018), the homogenous group of psychology students in our study also varied greatly in their conceptions of creativity. We found both quantitative and qualitative individual differences in how they valued originality and utility (Haaf & Rouder, 2017, 2018). Originality was usually important, but a few people seemed to regard original stimuli "not creative". Regarding utility, there were large individual differences too. Some participants seemed to value utility, a few rated any useful idea as "not creative", and some appeared to disregard it entirely. Moreover, the implicit values were negatively correlated (although in Study 2 we partly attributed this association to the stimulus correlation): the more participants valued originality, the less they valued utility when judging creativity. Future

[10] We conducted a Bayesian model comparison using Bridge sampling (not pre-registered; Meng & Wong, 1996) where we compare our original model to a model that has neither an overall nor a random utility effect on the drift rate. We found extreme evidence for our original model over the model without any utility effects. We report this analysis in the supplementary materials. We thank an anonymous reviewer for suggesting this analysis.

research should look to generalize these results to other populations or types of creativity judgments. For example, adolescents appear to place more value on originality and less on utility (Stevenson, 2022), so there may be a developmental trend. Also, judging the creativity of AUT ideas is interesting as the task is used so often in psychological assessment and educational assessment (e.g., Benedek et al., 2016). But, would the utility of an idea be considered more valuable in real-world situations, for example when judging the creativity of ideas to combat climate change?

After the CON-task, we asked participants to rate how important they thought originality and utility were for creativity. These self-reported relevance ratings more or less corresponded to their implicit values of originality and utility on the CON-task. For example, those who strongly and explicitly indicated that originality/innovativeness is important for creativity also tended to implicitly value originality in the CON-task. In addition, it appears that the CON-task provides an indirect way to tap into what people consider creative.

Similar to Caroff and Besançon (2008), our results also showed that the more participants took originality and utility into account on the CON-task, the more original and useful their responses to the AUT were. We cannot talk about causality, but it seems plausible that people's implicit values of originality and utility influence what ideas they produce and how they evaluate and select ideas during divergent thinking, like on the AUT. This coincides with findings that the more creative someone is, the better their creativity judgements are, whether its of their own or other's ideas (e.g., Benedek et al., 2016; Silvia, 2008).

Methodological Implications

To our knowledge, this is the first time a mathematical model such as the DDM has been applied to the evaluation phase of divergent thinking. The DDM is generally applied to timed two-choice decision tasks where there is clearly a correct answer, such as lexical decision tasks (e.g., Wagenmakers et al., 2008). Therefore, applying the DDM to the

CON-task, where there is no correct answer and we are basically assessing what factors influence people's tendency to decide if an idea is creative, is novel. Applying the DDM generally worked well on both studies, suggesting that the DDM might provide a reasonable account of the evaluation process in the AUT. A substantive interpretation of the evaluation process could be that when people judge creativity in the CON-task, they stochastically extract, accumulate, and integrate internal noisy signals about a stimulus' creativity, for example regarding the stimuli's originality and utility (for an interpretation regarding value-based decisions, see Milosavljevic et al., 2010).

Although overall we deemed the model fit to be acceptable, the model predicted slightly more right-skewed RT distributions at the individual level than there were in the data. This kind of misfit suggests that participants responded slower than the model predicted (but also note the predicted longer RTs in the upper quantiles of the RT distribution; see online supplementary materials). Since we explicitly instructed participants to respond as fast as possible, these longer RTs appear necessary for participants to evaluate creativity in the CON-task. The greater need for time is also reflected in a larger than usual boundary separation parameter which can be equated with the discrimination parameter in a 2PL model. It is directly influenced by the time participants have or take to respond (van der Maas et al., 2011; Wickelgren, 1977). Longer RTs therefore imply greater discriminatory power of the CON-task stimuli (van der Maas et al., 2011), which is desirable given our aim to study individual differences in creativity conceptions. However, the relatively long RTs might also suggest that creativity evaluation comprises relatively long, possibly sequential decision-making stages where one simple random walk as assumed in the DDM might be too simplistic. On the other hand, the DDM might still be a good enough approximation of the evaluation process as several more complex decision-making models for binary choices boil down to the DDM (Bogacz et al., 2006; van der Maas et al., 2011).

Regarding the role of RTs in our two studies, decision speed did not vary much across stimulus originality and utility ratings. Accordingly, there was no inverted u- or v-shaped

relationship between RTs and stimulus characteristics as one might have expected. At best, there was a tendency for stimuli with higher originality and utility ratings to be answered more slowly. A possible explanation is that the effects of originality and utility on the drift rate (i.e., people's tendency towards deciding "creative") were negatively correlated in both studies. This means that the more participants valued originality, the less they valued utility and vice versa. There appears to be a trade-off where people balance one dimension against the other when judging creativity, which could explain why stimuli with both higher originality and utility were processed more slowly.

The lacking influence of RTs begs the question of whether the DDM is not overly complex for the CON-task data from our two studies. For example, a hierarchical probit analysis[11] of our data resulted in similar conclusions. It would be interesting to examine whether RT effects are more apparent if participants are given less time (e.g., 3s versus the 9s in this study); especially since in the real-world people sometimes need to make split-second choices about which idea to pursue. Perhaps this would also lead to a better DDM model fit regarding the RT distribution. Also, instructions that emphasize speed more could affect the evaluation process, just as those of quantity or quality affect the ideation phase of divergent thinking (Said-Metwaly, Fernández-Castilla, Kyndt, & Van den Noortgate, 2020).

Conclusion

This paper demonstrates a new approach to studying the evaluation process during divergent thinking. Our novel application of the drift diffusion model provides a mathematical method to study how people decide what's creative. The results imply that it is important to take both originality and utility into account when examining how people judge creativity. Also, given how conceptions of creativity vary, our findings suggest that when assessing creativity using divergent thinking tasks like the AUT, managers or researchers should clarify what the rating criteria are to provide a fair playing field for all.

---

[11] Exploratory, not pre-registered, see online supplementary materials

Appendices

Appendix A

This Appendix contains the DDM specification. Model differences between the studies are mentioned where necessary.

Let $\boldsymbol{Y_{(ij)}}$ denote a response vector of the decision and response time $(X_{(ij)}, T_{(ij)})$ for the ith participant, $i = 1, ..., I$ in the jth trial (or stimulus), $j = 1, ..., J$ of the CON task. The bivariate data $\boldsymbol{Y_{(ij)}}$ is assumed to be distributed according to a Wiener distribution,

$$\boldsymbol{Y_{(ij)}} \sim \text{Wiener}(\alpha_{(ij)}, \beta_{(ij)}, \tau_{(ij)}, \delta_{(ij)}),$$

with the four model parameters boundary separation α, bias β, non-decision-time τ, and drift rate δ. The Wiener distribution is a joint density function of deciding whether a use is creative or not, $X_{(ij)}$, at time $T_{(ij)}$ (Vandekerckhove et al., 2011).

The double index notation suggests that, in principle, the four parameters may differ across participants, as well as across trials. To reduce model complexity, we constrain the model in several ways. First, we treat all parameters as constant across trials. Second, at the participant level, we allow participants to differ in three out of four parameters, α, β, and δ. The non-decision time parameter τ is constrained to be constant across trials as well as across participants (i.e., $\tau_{(ij)} = \tau$), because interpreting random effects for the non-decision time parameter has shown to be problematic (Singmann, 2018b). We treat differences across individuals as random effects, assuming that participants are a sample from a population distribution.

Our main focus of interest is on the drift rate parameter δ. It is the only model parameter assumed to be influenced by stimulus characteristics because the remaining

parameters are already set before the decision-making of whether something is creative or not takes place (e.g., Vandekerckhove et al., 2011). To examine the influence of originality and utility when evaluating creativity, we regress δ on the originality and utility ratings of the stimuli. We include random intercepts as well as random slopes to take interindividual variation into account. Furthermore, because the response times and proportions of "creative" responses have shown to vary considerably across the 64 task stimuli, we also include random intercepts pertaining to the stimuli. Mathematically, we express the above described as follows,

$$\delta_{(ij)} = \theta_{\delta(i)} + \phi_{\delta(j)} + \theta_{OR(i)} z_{OR(j)} + \theta_{UT(i)} z_{UT(j)}.$$

The parameters $\theta_{\delta(i)}$, $\theta_{OR(i)}$, $\theta_{UT(i)}$, and $\phi_{\delta(j)}$ reflect the drift rate decomposition. Specifically, $\theta_{\delta(i)}$ denotes the drift rate intercept, $\theta_{OR(i)}$ the originality effect, and $\theta_{UT(i)}$ the utility effect of individual i. Furthermore, $\phi_{\delta(j)}$ is stimulus j's individual deviation from the drift rate intercept. Lastly, $z_{OR(j)}$ and $z_{UT(j)}$ refer to z-scores of the originality and utility ratings of stimulus j.

In Study 1, the boundary separation parameter has random and fixed effects and the bias parameter is fixed. In Study 2, both the boundary separation and the bias parameter are allowed to vary across individuals. All random effects pertaining to the individuals are correlated in both studies. As such, we assume that the random effects are drawn from the same multivariate normal distribution with variance-covariance matrix $\mathbf{\Sigma}$, i.e.,

$$\begin{bmatrix} \theta_{\delta(i)} \\ \theta_{OR(i)} \\ \theta_{UT(i)} \\ \alpha_{(i)} \end{bmatrix} \sim \text{Multivariate-Normal} \left(\begin{bmatrix} \mu_{\delta} \\ \mu_{\theta_{OR}} \\ \mu_{\theta_{UT}} \\ \mu_{\alpha} \end{bmatrix}, \mathbf{\Sigma} \right),$$

in Study 1, and

$$\begin{bmatrix} \theta_{\delta(i)} \\ \theta_{OR(i)} \\ \theta_{UT(i)} \\ \alpha_{(i)} \\ \beta_{(i)} \end{bmatrix} \sim \text{Multivariate-Normal} \left(\begin{bmatrix} \mu_{\delta} \\ \mu_{\theta_{OR}} \\ \mu_{\theta_{UT}} \\ \mu_{\alpha} \\ \mu_{\beta} \end{bmatrix}, \boldsymbol{\Sigma} \right),$$

in Study 2.

$\boldsymbol{\Sigma}$ is further defined below. The random stimulus effects are orthogonal to the random effects concering the individuals. They are also assumed to be randomly sampled from a population distribution (of stimuli),

$$\phi_{\delta(j)} \sim \text{Normal}(0, \sigma_{\delta_\phi}),$$

where 0 is the mean and σ_{δ_ϕ} is the standard deviation.

We need to specify priors for all fixed and random effects parameters as well as for the correlations among the random effects parameters.

Prior specification Study 1. We use a standard normal prior for the originality and utility effects on the drift rate,

$$\theta_{OR}, \theta_{UT} \sim \text{Normal}(0, 1),$$

.

For the remaining fixed effects we use the following weakly informative priors:

$$\theta_\delta \sim \text{Normal}(0, 1)$$

$$\mu_\beta \sim \text{Beta}(1.3, 1.3)$$

$$\mu_\alpha \sim \text{Normal}^+(0, 2)$$

$$\tau \sim \text{Uniform}(0, 0.3).$$

These prior distributions restrict the parameters to a plausible range. The range of the a-priori bias parameter is from 0 to 1 and the boundary separation is restricted to be positive. The non-decision time generally needs to be smaller than the RTs. We therefore use 0.3 seconds, the minimally required response time (see exclusion criterion II.), for the prior on τ. Note that due to model convergence issues, we increase the upper bound by one millisecond to 0.301 which is the minimum RT in the data.

For all variability parameters we use the following prior,

$$\sigma_{\delta_\nu}, \sigma_{\delta_\phi}, \sigma_{OR}, \sigma_{UT}, \sigma_\alpha \sim \text{Normal}^+(0, 0.3).$$

Lastly, we place a prior on the random effects correlations concerning the individuals. The variance-covariance matrix $\mathbf{\Sigma}$ needs to be decomposed such that we can specify a prior for the correlations only. We refer to the matrix containing the random effects correlations as $\mathbf{P}$. Specifically, $\mathbf{\Sigma}$ can be rewritten as $\mathbf{\Phi P \Phi}$, whereby $\mathbf{\Phi}$ is a 4x4 matrix with only the variability parameters on the diagonal, $\mathbf{\Phi} = \text{diag}(\sigma_{\delta_\nu}, \sigma_{OR}, \sigma_{UT}, \sigma_\alpha)$, and $\mathbf{P}$ is a 4x4 correlation matrix. For example, the correlation between the random originality and utility effects is expressed as $\rho_{\sigma_{OR}\sigma_{UT}}$. We place a Lewandowski-Kurowicka-Joe (LKJ) prior with the shape 3 on $\mathbf{P}$ (Lewandowski, Kurowicka, & Joe, 2009),

$$\mathbf{P} \sim \text{LKJ}(3).$$

This prior restricts the correlations to the range -1 to 1, makes it a proper correlation matrix, and places most prior mass around 0.

Prior specification Study 2. We use informative, truncated prior distributions for the parameters where we expect a positive effect (i.e., the originality and utility effects on the drift rate),

$$\mu_{\theta_{OR}}, \mu_{\theta_{UT}} \sim \text{Normal}^{+}(0, 0.2),$$

where 0 is the mean and 0.2 the standard deviation. This truncated prior distribution is informed by previous research that, overall, people take into account both originality and utility when they evaluate creative ideas. It is also informed by data. In Study 1's dataset, the presence of effects of both stimulus originality and utility on the drift rate are detectable using this prior.

For the remaining fixed effects we use the following weakly informative priors:

$$\mu_{\theta_\delta} \sim \text{Normal}(0, 1)$$

$$\mu_\beta \sim \text{Beta}(1, 1)$$

$$\mu_\alpha \sim \text{Normal}^{+}(0, 2)$$

$$\tau \sim \text{Uniform}(0, 0.3).$$

These prior distributions again restrict the parameters to a plausible range. Note that to successfully estimate the model, the upper bound of the uniform prior on τ is again set to 0.301 (instead of 0.300), the minimum response time in Study 2.

For all variability parameters we use the following prior,

$$\sigma_{\delta_\nu}, \sigma_{\delta_\phi}, \sigma_{OR}, \sigma_{UT}, \sigma_\beta, \sigma_\alpha \sim \text{Normal}^+(0, 0.3).$$

This prior is again informed by previous analyses on Study 1's dataset. Lastly, we need to place a prior on the random effects correlations concerning the individuals. Here, $\mathbf{\Sigma}$ can again be rewritten as $\mathbf{\Phi P \Phi}$, whereby $\mathbf{\Phi}$ is a 5x5 matrix with only the variability parameters on the diagonal, $\mathbf{\Phi} = \text{diag}(\sigma_{\delta_\nu}, \sigma_{OR}, \sigma_{UT}, \sigma_\alpha, \sigma_\beta)$, and $\mathbf{P}$ is a 5x5 correlation matrix. We again place a Lewandowski-Kurowicka-Joe (LKJ) prior with the shape parameter 3 on $\mathbf{P}$ (Lewandowski et al., 2009),

$$\mathbf{P} \sim \text{LKJ}(3).$$

Appendix B

This Appendix contains the data cleaning procedure that we applied in Study 1 and Study 2.

Study 1.

CON-task. Before analyzing the data, we employed the following exclusion criteria. The full dataset comprised 18984 trials from 299 participants. First, we excluded data from all participants who gave the same response (either "creative" or "not creative") in at least 57/64 ($\approx 90\%$) of the trials. This step removed data from 2 individuals. We then removed the first two trials to account for the fact that participants needed time to get acquainted with the task (= 594 trials). We also excluded all trials with response times greater than 6 seconds and less than 0.3 seconds to exclude unreasonably fast and slow responses. In this step, 337 trials were excluded. Finally, we excluded data from individuals with fewer than 47 ($\approx 3/4$) remaining trials. This last step removed data from 4 participants.

AUT. Out of the participants who were included in the CON-task analysis, only 1 did not complete the AUT which left us with 292 participants for the AUT analysis. However, 6 participants did not submit responses to one of the two AUT objects. To clean the AUT data, we first removed within-participant duplicates (e.g., from the responses "toy 1", "toy 2", "toy 3" for the object "brick", we only kept the first response). We then removed data from all participants with less than 90 percent valid responses ($n = 66$). We treated a response as "invalid" if at least one rater had scored it as such (i.e., a rating of "0"). Examples are responses where participants responded with associations rather than uses (e.g., "rectangular" as response for the object "brick"). Finally, we removed all responses that both raters had scored as invalid before computing the performance indicators. As performance indicators, we used the mean originality and mean utility score across raters, objects, and responses.

Study 2.

CON-task. The full dataset comprised 10972 trials and 172 participants. First, we excluded 8 participants for giving the same response in at least 57/64 ($\approx 90\%$) of the trials. We then removed the first two trials for each participant (= 328 trials) and all trials with RTs greater than 6 seconds and less than 0.3 seconds (= 529 trials). Finally, data from 12 participants were excluded because they had fewer than 47 ($\approx 3/4$) remaining trials. The sample used to estimate the DDM comprised 152 participants and 9291 trials. Although we pre-registered to include only data from Dutch native speakers, we retrospectively decided not to exclude data from non-Dutch native speakers as long as they were able to read and respond fluently in Dutch to be as inclusive as possible. Thus, we included all participants who chose to do the Dutch (rather than English) version of the experiment. Participants were required to read an instruction in Dutch in order to do the experiment in Dutch. If they chose the Dutch version, we considered them sufficiently fluent to read and respond in Dutch for our study's purposes where only a few simple phrases had to be read or written in Dutch.

AUT. In Study 2, 14 participants who were included in the CON-task analyses did not complete the AUT. Again a few participants ($n = 3$) only submitted responses for one of the two objects. We again cleaned the AUT data by removing all within-participant duplicates as well as data from all participants with less than 90 percent responses that were scored as valid by both raters. Additionally, we again removed all responses that both raters had scored as invalid ($n = 12$). As performance indicators, we again used the mean originality and mean utility score across raters, objects and responses.

Appendix C

This Appendix shows the response time distributions of the cleaned data in Study 1 and Study 2.

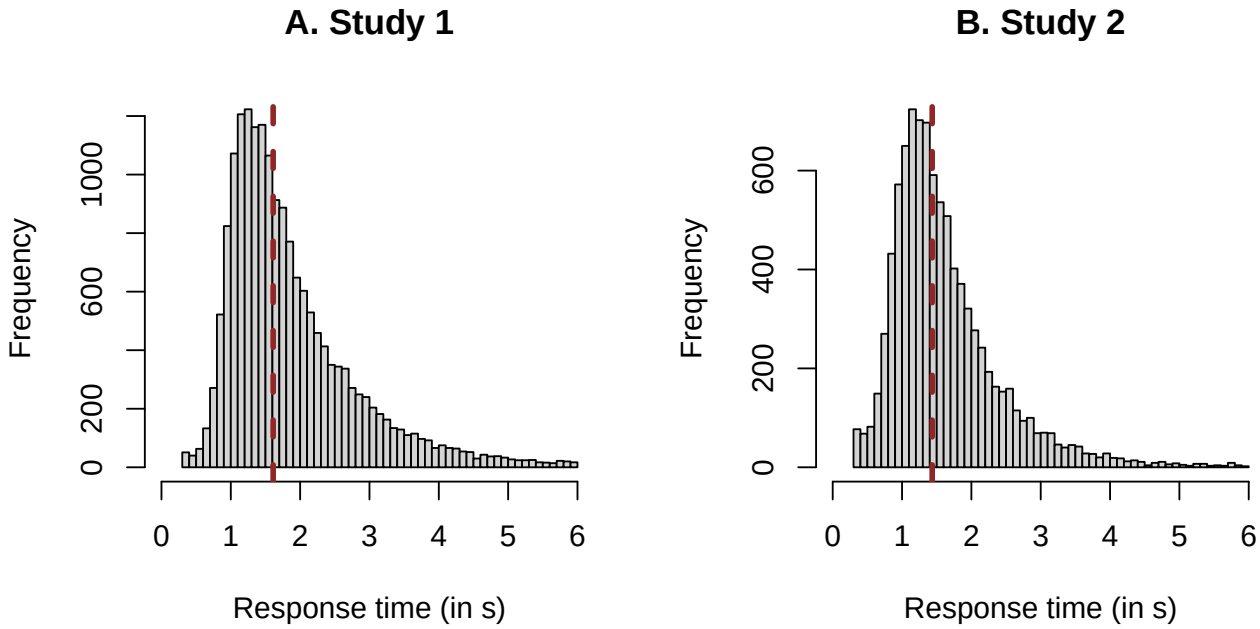

Figure A1. A and B show histograms of the response times in Study 1 and Study 2, respectively. The dotted lines denote the response time medians.

(#fig:RT distribution plot)

Appendix D

Table A1

Posterior mean, standard deviation, and 95% credible interval for the correlations among random effects parameters

	Study 1				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
$\rho_{\sigma_{\theta_{OR}}\sigma_{\theta_{UT}}}$	-0.44	0.08	-0.6	-0.27	-0.34	0.13	-0.59	-0.09
$\rho_{\sigma_{\delta_\nu}\sigma_{\theta_{OR}}}$	-0.16	0.07	-0.3	-0.02	-0.03	0.11	-0.24	0.18
$\rho_{\sigma_{\delta_\nu}\sigma_\alpha}$	-0.1	0.07	-0.24	0.03	0.03	0.09	-0.16	0.21
$\rho_{\sigma_{\theta_{UT}}\sigma_\alpha}$	-0.06	0.08	-0.21	0.1	-0.21	0.12	-0.43	0.02
$\rho_{\sigma_{\delta_\nu}\sigma_{\theta_{UT}}}$	0.07	0.08	-0.09	0.22	-0.09	0.11	-0.30	0.11
$\rho_{\sigma_{\theta_{OR}}\sigma_\alpha}$	0.31	0.07	0.16	0.45	0.24	0.11	0.01	0.45
$\rho_{\sigma_{\delta_\nu}\sigma_\beta}$	-	-	-	-	-0.45	0.10	-0.63	-0.24
$\rho_{\sigma_{\theta_{OR}}\sigma_\beta}$	-	-	-	-	-0.17	0.14	-0.44	0.12
$\rho_{\sigma_{\theta_{UT}}\sigma_\beta}$	-	-	-	-	0.14	0.14	-0.14	0.41
$\rho_{\sigma_\alpha\sigma_\beta}$	-	-	-	-	-0.42	0.10	-0.61	-0.21

Note. SD = standard deviation; LB = lower bound; UB = upper bound.

References

al., A. S. et mult. (2019). *DescTools: Tools for descriptive statistics*. Retrieved from https://cran.r-project.org/package=DescTools

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2020). *Magrittr: A forward-pipe operator for r*. Retrieved from https://CRAN.R-project.org/package=magrittr

Barchard, K. A. (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. *Psychological Methods*, *17*(2), 294–308. https://doi.org/10.1037/a0023351

Barron, F. (1955). The disposition toward originality. *The Journal of Abnormal and Social Psychology*, *51*(3), 478–485. https://doi.org/10.1037/h0048073

Basadur, M. (1995). Optimal ideation-evaluation ratios. *Creativity Research Journal*, *8*(1), 63–75. https://doi.org/10.1207/s15326934crj0801_5

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Benedek, M., Fink, A., & Neubauer, A. C. (2006). Enhancement of ideational fluency by means of computer-based training. *Creativity Research Journal*, *18*(3), 317–328. https://doi.org/10.1207/s15326934crj1803_7

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., &

Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity*, *21*, 75–84. https://doi.org/10.1016/j.tsc.2016.05.007

Berkelaar, M., & others. (2020). Retrieved from https://CRAN.R-project.org/package=lpSolve

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd edition). New York, NY: Routledge.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Caroff, X., & Besançon, M. (2008). Variability of creativity judgments. *Learning and Individual Differences*, *18*(4), 367–371. https://doi.org/10.1016/j.lindif.2008.04.001

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers'*

perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. Workforce. Washington, DC: Partnership for 21st Century Skills.

Davies, M. (2008). *The corpus of contemporary american english (coca)*. Available online at https://www.english-corpora.org/coca/.

de Buisonjé, D. R., Ritter, S. M., de Bruin, S., ter Horst, J. M.-L., & Meeldijk, A. (2017). Facilitating creative idea selection: The combined effects of self-affirmation, promotion focus and positive affect. *Creativity Research Journal*, *29*(2), 174–181. https://doi.org/10.1080/10400419.2017.1303308

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204–223.

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

Donzallaz, M. C., Haaf, J. M., & Stevenson, C. (2022, July 1). Creative or not? Hierarchical diffusion modeling of the creative evaluation process. https://doi.org/10.17605/OSF.IO/73C2D

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, *72*(1), 28–36. https://doi.org/10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Efron, B., & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American*, *236*(5), 119–127. https://doi.org/10.1038/scientificamerican0577-119

Forthmann, B., Jendryczko, D., Scharfen, J., Kleinkorres, R., Benedek, M., & Holling, H. (2019). Creative ideation, broad retrieval ability, and processing speed: A confirmatory study of nested cognitive abilities. *Intelligence*, *75*, 59–72. https://doi.org/10.1016/j.intell.2019.04.006

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *J. R. Stat. Soc. A*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gamer, M., Lemon, J., & <puspendra.pusp22@gmail.com>, I. F. P. S. (2019). *Irr: Various coefficients of interrater reliability and agreement*. Retrieved from https://CRAN.R-project.org/package=irr

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Heidelberg: Springer-Verlag.

Grohman, M., Wodniecka, Z., & Kłusak, M. (2006). Divergent Thinking and Evaluation Skills: Do They Always Go Together? *The Journal of Creative Behavior*, *40*(2), 125–145. https://doi.org/10.1002/j.2162-6057.2006.tb01269.x

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

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models. *Psychological Methods*, *22*, 779–798. https://doi.org/10.1037/met0000156

Haaf, J. M., & Rouder, J. N. (2018). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-018-1522-x

Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, *61*(1), 569–598. https://doi.org/10.1146/annurev.psych.093008.100416

Herman, A., & Reiter-Palmon, R. (2011). The effect of regulatory focus on idea generation and idea evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(1), 13–20. https://doi.org/10.1037/a0018587

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

IBM. (2010). *Capitalizing on complexity.* IBM Global CEO Study.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. Retrieved from https://www.jstatsoft.org/v38/i08/

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krypotos, A.-M., Beckers, T., Kindt, M., & Wagenmakers, E.-J. (2015). A Bayesian hierarchical diffusion model decomposition of performance in Approach–Avoidance

Tasks. *Cognition and Emotion*, *29*(8), 1424–1444. https://doi.org/10.1080/02699931.2014.985635

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Loewenstein, J., & Mueller, J. (2016). Implicit theories of creative ideas: How culture guides creativity assessments. *Academy of Management Discoveries*, *2*(4), 320–348. https://doi.org/10.5465/amd.2014.0147

Mastria, S., Agnoli, S., & Corazza, G. E. (2019). How does emotion influence the creativity evaluation of exogenous alternative ideas? *PLOS ONE*, *14*(7), e0219298. https://doi.org/10.1371/journal.pone.0219298

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

Meng, X.-L., & Wong, W. H. (1996). Simulation ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The Drift

Diffusion Model Can Account for the Accuracy and Reaction Time of Value-Based Choices Under High and Low Time Pressure. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1901533

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Mueller, J., Melwani, S., Loewenstein, J., & Deal, J. J. (2018). Reframing the decision-makers' dilemma: Towards a social context model of creative idea recognition. *Academy of Management Journal*, *61*(1), 94–110. https://doi.org/10.5465/amj.2013.0887

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

Puente-Diaz, R., Cavazos-Arroyo, J., & Puerta-Sierra, L. (2021). Idea Generation, Selection, and Evaluation: A Metacognitive Approach. *The Journal of Creative Behavior*, jocb.505. https://doi.org/10.1002/jocb.505

Randolph, M. (2019). *That will never work: The birth of Netflix and the amazing life of an idea*. New York, NY: Hachette Book Group.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295x.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. https://doi.org/10.3758/BF03196283

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279. https://doi.org/10.1037/dec0000030

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review*, *111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2018). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2010). The selection of creative ideas after individual idea generation: Choosing between creativity and impact. *British Journal of Psychology*, *101*(1), 47–68. https://doi.org/10.1348/000712609X414204

Ritter, S. M., van Baaren, R. B., & Dijksterhuis, A. (2012). Creativity: The role of unconscious processes in idea generation and idea selection. *Thinking Skills and Creativity*, *7*(1), 21–27. https://doi.org/10.1016/j.tsc.2011.12.002

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, *24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., & Basadur, M. (1993). Assessing ideational and evaluative skills and creative styles and attitudes. *Creativity and Innovation Management*, *2*(3). https://doi.org/10.1109/iemc.1990.201291

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, *15*(5), 537–546. https://doi.org/10.1016/0191-8869(93)90337-3

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

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences*, *13*(3), 295–302. https://doi.org/10.1016/0191-8869(92)90105-x

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. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139–146. https://doi.org/10.1037/1931-3896.2.3.139

Simonton, D. K. (2018). Defining Creativity: Don't We Also Need to Define What Is *Not* Creative? *The Journal of Creative Behavior*, *52*(1), 80–90. https://doi.org/10.1002/jocb.137

Singmann, H. (2018a, January 7). Diffusion/Wiener Model Analysis with brms – Part II: Model Diagnostics and Model Fit. Retrieved from http://singmann.org/wiener-model-analysis-with-brms-part-ii/

Singmann, H. (2018b, September 6). Diffusion/Wiener Model Analysis with brms – Part III: Hypothesis Tests of Parameter Estimates. Retrieved from http://singmann.org/wiener-model-analysis-with-brms-part-iii/

Stan Development Team. (2019a). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Stan Development Team. (2019b). StanHeaders: Headers for the R interface to Stan. Retrieved from http://mc-stan.org/

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

Sternberg, R. J., & Lubart, T. I. (1996). Investing in creativity, *51*(7), 677–688.

Stevenson, C. E. (2022). Creative or not: The originality-utility tradeoff in divergent thinking [conference presentation]. Society for the Neuroscience of Creativity (SfNC) annual meeting May 12-13th 2022.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 702–716. https://doi.org/10.3758/BF03206552

Tuerlinckx, F., & Boeck, P. D. (2005). Two interpretations of the discrimination parameter.

Psychometrika, *70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356. https://doi.org/10.1037/a0022749

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020, January 16). Rank-normalization, folding, and localization: An improved $\widehat{}R{}$ for assessing convergence of MCMC. Retrieved February 4, 2020, from http://arxiv.org/abs/1903.08008

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener package: An R package providing distribution functions for the Wiener diffusion model. *The R Journal*, *6*(1), 49. https://doi.org/10.32614/RJ-2014-005

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671. https://doi.org/10.1080/09541440802205067

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Wei, T., & Simko, V. (2017). *R package "corrplot": Visualization of a correlation matrix*. Retrieved from https://github.com/taiyun/corrplot

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–65. https://doi.org/10.1016/0001-6918(77)90012-9

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wilke, C. O. (2019). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot