Supplementary Materials: 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 This paper has not been peer reviewed.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

Supplementary materials of the paper "Creative or Not? Hierarchical Diffusion Modeling of the Creative Evaluation Process".

Correspondence concerning this article should be addressed to Claire E. Stevenson, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam. E-mail: c.e.stevenson@uva.nl

Supplementary Materials: Creative or Not? Hierarchical Diffusion Modeling of the Creative Evaluation Process

Posterior predictive checks

Here we describe how we assessed the fit of the drift diffusion model (DDM) applied in Study 1 and 2 using posterior predictive checks. We followed the procedure described by Singmann (2018)]. The overall goal of the check was to examine whether the model was able to adequately describe the observed data. To this end, we obtained 500 sampled datasets from the posterior predictive distribution in both Study 1 and Study 2. First, we examined whether the model could reproduce the general response time (RT) and response proportion pattern observed in the data. We calculated three summary statistics per participant and sampled dataset. The summary statistics were (1) the proportion of "creative" responses, (2) the median RT for "creative" responses, and (3) the median RT for "not creative" responses. We then summarized these statistics further by taking the median and additional quantiles across datasets per participant. Lastly, we took the mean over all participants for each statistic and calculated the three statistics also for the observed data. Figure S1 shows the three summary statistics of the predictions (in black and grey; including credible intervals) and of the data (in red) for Study 1 and Figure S2 shows the same for Study 2. Overall, the models were able to adequately describe the general patterns in the data.

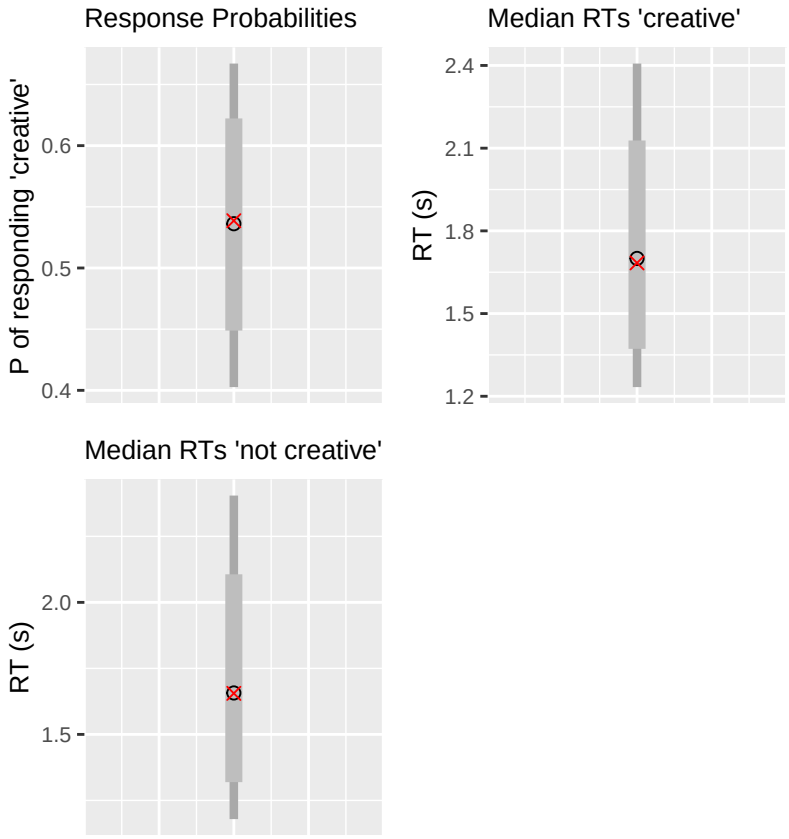

Figure S1. Posterior predictive check of Study 1. The inner vertical lines show the 80% CrIs and the outer vertical lines the 95% CrIs. The red cross denotes the observed and the black circle the predicted statistics. The model could reproduce the three summary statistics quite accurately.

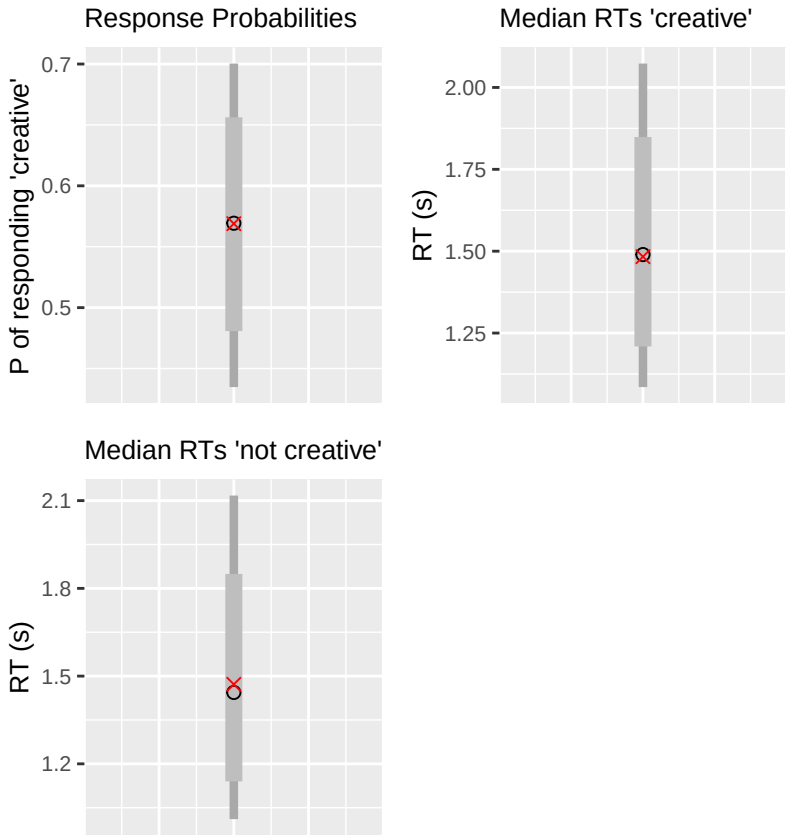

Figure S2. Posterior predictive check of Study 2. The inner vertical lines show the 80% CrIs and the outer vertical lines the 95% CrIs. The red cross denotes the observed and the black circle the predicted statistics. The model could reproduce the three summary statistics quite accurately.

Second, we computed the coverage probabilities of the three summary statistics across participants, (Singmann, 2018). For each of the statistics and for different credible intervals (CrIs), we calculated whether the three observed statistics were covered by the corresponding CrI. The coverage probabilities should be at least the width of the CrI (e.g., 50% for 50% CrI). They are shown in Table S1 for Study 1 and in Table S2 for Study 2. In both datasets, they corresponded with the width of the CrIs and even above. For several measures, the coverage probability was even 1 for the 95% and the 99% CrIs in both datasets/studies.

Table S1

Study 1: Coverage probabilities for the three summary statistics across participants and the 50%, 80%, 95%, and 99% CrIs

Statistic	50% CrI	80% CrI	95% CrI	99% CrI
Median RT 'not creative'	0.724	0.942	0.993	0.993
Median RT 'creative'	0.686	0.956	1.000	1.000
Proportion 'creative'	0.959	1.000	1.000	1.000

Table S2

Study 2: Coverage probabilities for the three summary statistics across participants and the 50%, 80%, 95%, and 99% CrIs

Statistic	50% CrI	80% CrI	95% CrI	99% CrI
Median RT 'not creative'	0.789	0.974	1.000	1.000
Median RT 'creative'	0.809	0.987	1.000	1.000
Proportion 'creative'	0.947	0.980	0.993	0.993

Third, we assessed the model fit by inspecting more RT quantiles than just the median as well as the "creative" response proportions, again closely following Singmann (2018). Specifically, we examined the observed and predicted RT quantiles (i.e., the 10th, 25th, 75th, and 90th) across participants. We first computed the quantiles for each sample of the posterior predictive distribution and then aggregated them. To assess the extent to which the observed and predicted quantiles matched, we calculated the concordance correlation coefficient for each quantile (CCC; e.g., Barchard, 2012). The CCC indicates the extent of absolute agreement between two values and ranges from -1 to 1, whereby CCC = 0 stands for no agreement, CCC = 1 for perfect agreement, and CCC = -1 for perfect disagreement. Figure S3 and Figure S4 show a QQ-plot for each quantile and for each response option. In

general and across studies, the fit was slightly better for the "creative" responses compared to the "not creative" responses. The model fit was best for the medians and worst for the 10th quantiles. At the 10th and 25th quantiles, the predicted RTs were smaller than the observed ones which could be a sign of shrinkage. At the 75th quantiles, on the other hand, the model predicted slightly slower RTs than were observed. This misfit suggests that the model predicted a more right-skewed response time distribution with fatter tails than observed (also see Figure S6). Apart from the predicted RT patterns, we also examined the observed vs. predicted response proportions across participants. As can be seen in Figure S5, the model was able to reproduce the proportion of "creative" responses quite well apart from a few outliers. Overall, apart from some misfit in the outer quantiles of the RT distribution, the model could reproduce the data quite accurately and appeared to provide an acceptable account of the data.

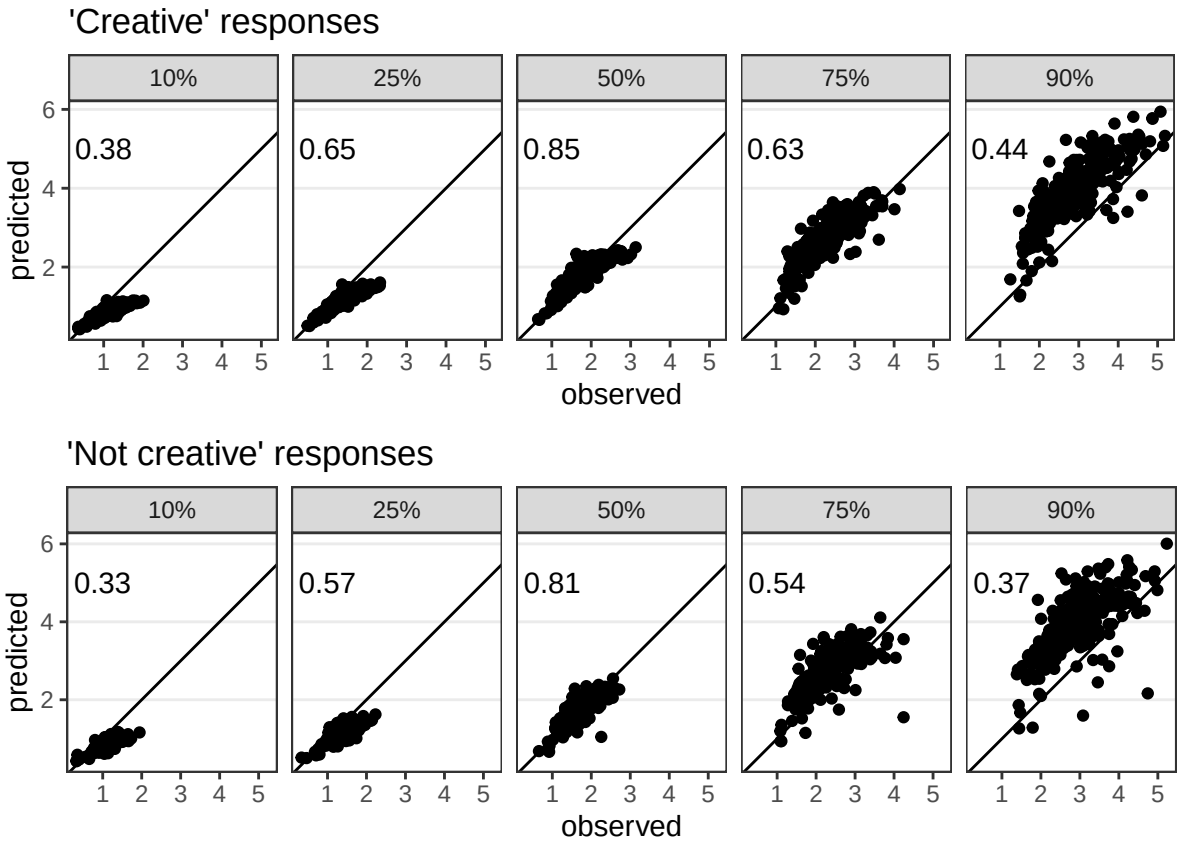

Figure S3. Posterior predictive check of Study 1. Each dot represents a participant. The model fit was slightly better for the upper responses compared to the lower responses. The fit was best for the medians and worst for the 10th quantiles. At the 10th and 25th quantiles, the predicted RTs were smaller than the observed ones. At the 75th quantiles the model predicted slightly slower RTs than were observed, which could indicate some bias in the model. Apart from that, the model fits well.

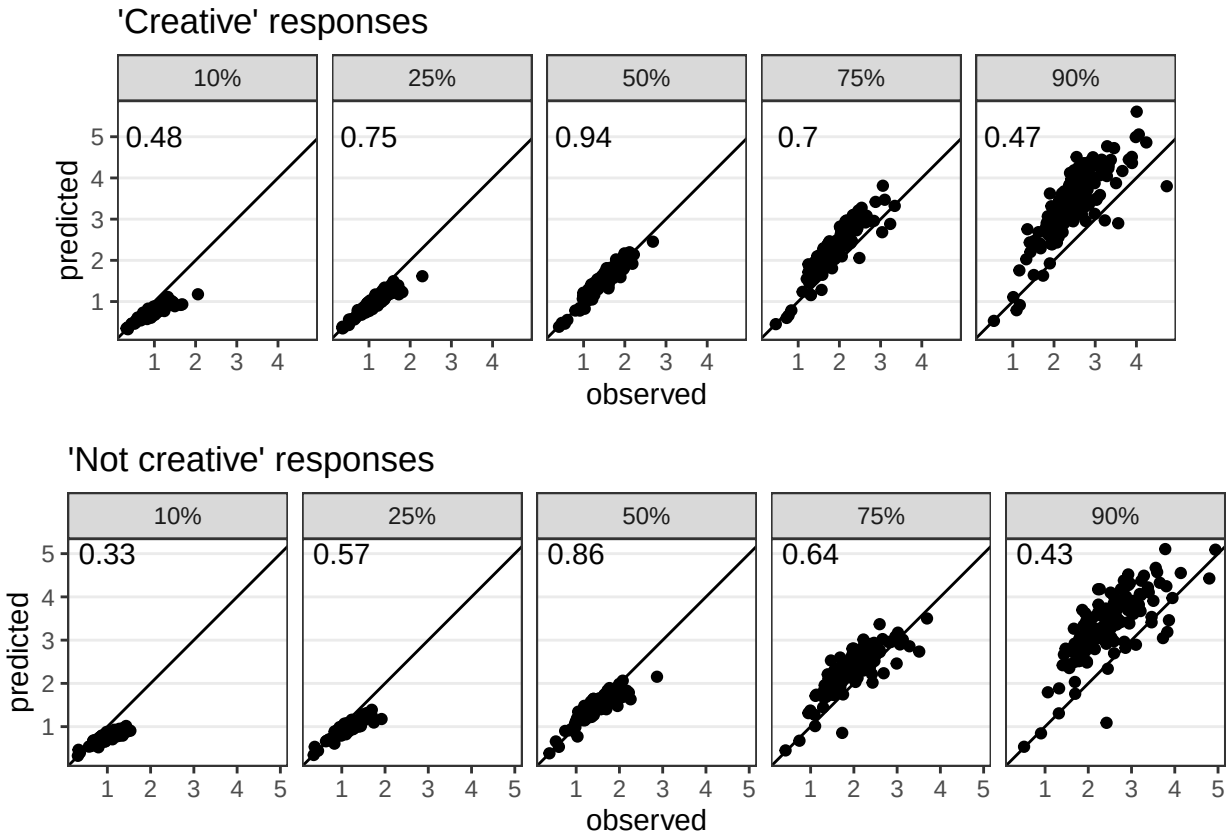

Figure S4. Posterior predictive check of Study 2. Each dot represents a participant. The model fit was slightly better for the upper responses compared to the lower responses. The fit was best for the medians and worst for the 10th quantiles. At the 10th and 25th quantiles, the predicted RTs were smaller than the observed ones. At the 75th quantiles the model predicted slightly slower RTs than were observed, which could indicate some bias in the model. Apart from that, the model fits well.

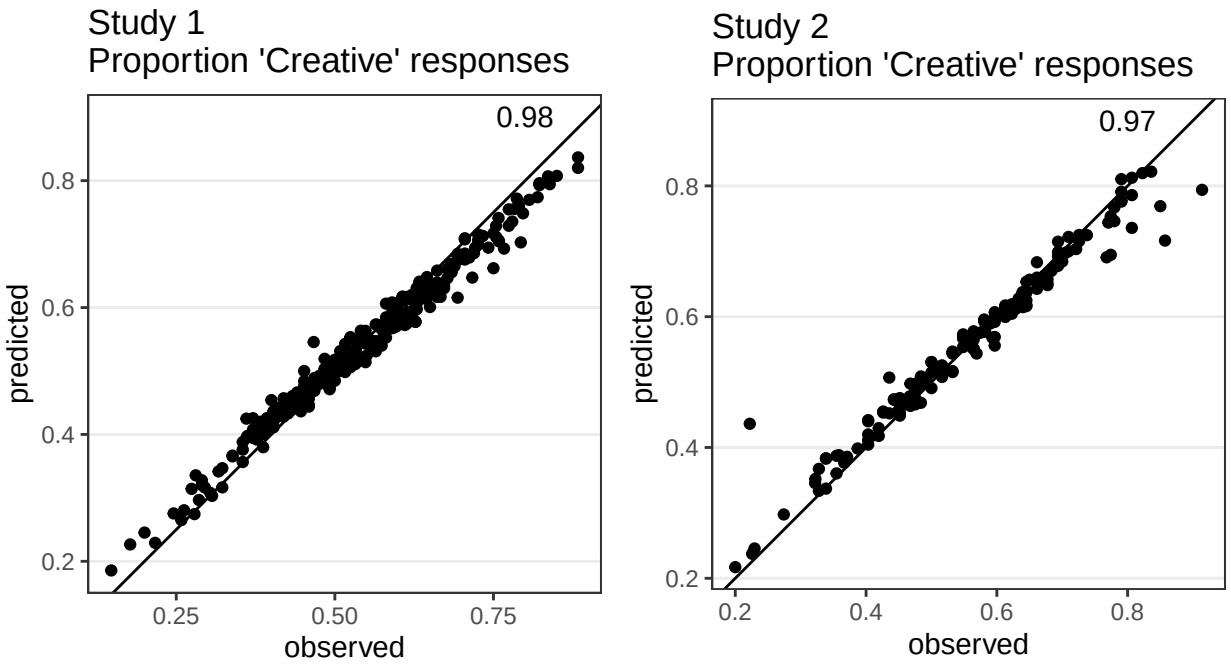

Figure S5. Posterior predictive check of the proportion of 'creative' responses in Study 1 and Study 2. Each dot represents a participant.

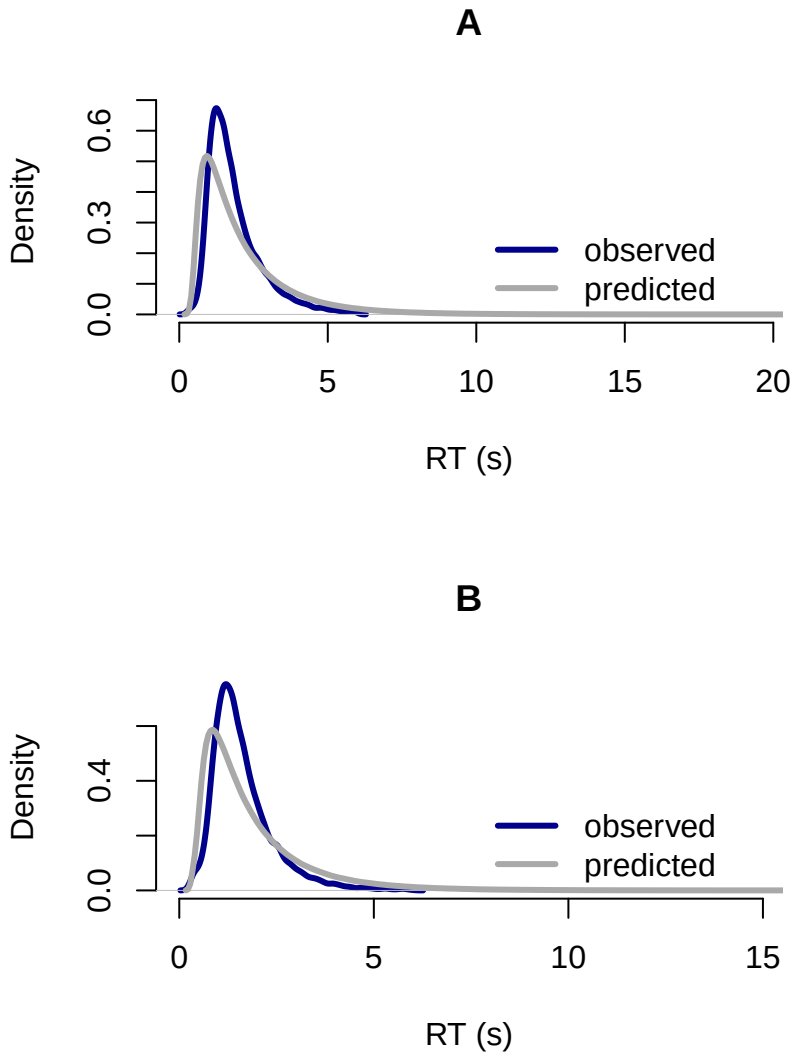

Figure S6. Observed and predicted response time distributions in Study 1 (A) and 2 (B). The upper limit of the y-axis is determined by 1/2 of the maximum RT in the posterior predictive distribution samples.

Model Estimation Including Participants Who Were Excluded Based on Too Few Trials

We repeated the key analyses including the participants that we had excluded based on too few trials (<47) to examine whether we would have arrived at the same conclusions had we included them. As shown in Table S3, S4, and S5, the model estimation results differed only slightly.

The posterior means of the originality and utility effects on the drift rate were identical for the utility effect and differed only slightly for the originality effect. Furthermore, the posterior mean of the random effects correlation between the originality and utility effects was only somewhat smaller.

Table S3

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

	Study 2 (excl.)				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
μ_δ	0.14	0.05	0.04	0.25	0.15	0.05	0.05	0.26
μ_β	0.50	0.01	0.48	0.51	0.49	0.01	0.48	0.50
$\mu_{\theta_{OR}}$	0.38	0.05	0.28	0.49	0.40	0.05	0.30	0.51
$\mu_{\theta_{UT}}$	0.10	0.05	0.01	0.20	0.10	0.05	0.01	0.21
μ_α	2.63	0.05	2.53	2.72	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 S4

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 2 (excl.)				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
σ_{δ_ϕ}	0.32	0.03	0.27	0.39	0.33	0.03	0.27	0.40
σ_{δ_υ}	0.37	0.02	0.32	0.42	0.36	0.03	0.32	0.42
$\sigma_{\theta_{OR}}$	0.20	0.02	0.17	0.24	0.20	0.02	0.16	0.24
$\sigma_{\theta_{UT}}$	0.20	0.02	0.16	0.23	0.20	0.02	0.17	0.24
σ_α	0.60	0.03	0.54	0.67	0.46	0.03	0.41	0.52
σ_β	0.05	0.01	0.04	0.06	0.05	0.01	0.04	0.06

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

Table S5

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

	Study 2 (excl.)				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
$\rho_{\sigma_{\delta_\nu}\sigma_\beta}$	-0.51	0.09	-0.68	-0.32	-0.45	0.10	-0.63	-0.24
$\rho_{\sigma_\alpha\sigma_\beta}$	-0.45	0.10	-0.63	-0.25	-0.42	0.10	-0.61	-0.21
$\rho_{\sigma_{\theta_{OR}}\sigma_{\theta_{UT}}}$	-0.31	0.13	-0.56	-0.05	-0.34	0.13	-0.59	-0.09
$\rho_{\sigma_{\theta_{OR}}\sigma_\beta}$	-0.27	0.14	-0.53	0.01	-0.17	0.14	-0.44	0.12
$\rho_{\sigma_{\theta_{UT}}\sigma_\alpha}$	-0.10	0.12	-0.35	0.14	-0.21	0.12	-0.43	0.02
$\rho_{\sigma_{\delta_\nu}\sigma_{\theta_{UT}}}$	-0.08	0.10	-0.28	0.13	-0.09	0.11	-0.30	0.11
$\rho_{\sigma_{\delta_\nu}\sigma_{\theta_{OR}}}$	-0.01	0.11	-0.21	0.20	-0.03	0.11	-0.24	0.18
$\rho_{\sigma_{\delta_\nu}\sigma_\alpha}$	0.06	0.10	-0.13	0.25	0.03	0.09	-0.16	0.21
$\rho_{\sigma_{\theta_{UT}}\sigma_\beta}$	0.11	0.14	-0.17	0.37	0.14	0.14	-0.14	0.41
$\rho_{\sigma_{\theta_{OR}}\sigma_\alpha}$	0.34	0.12	0.09	0.55	0.24	0.11	0.01	0.45

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

Estimation With Uncorrelated Stimuli Set

Study 2

In the original analysis, we found a substantial correlation between the originality and utility effects, suggesting that the more individuals take originality into account when they evaluate creativity, the less they take utility into account and vice versa. However, the originality and utility ratings were negatively correlated (r = -0.61). The correlation between the originality and utility effects may therefore be a function of the stimuli.

To assess whether this is indeed the case, we re-estimated the model based on an uncorrelated set of stimuli and comparing results. We excluded all stimuli with an originality rating below 1.5 and a utility rating above -1 or a utility rating below 1.5. These criteria led us to exclude 20 stimuli, reducing the stimulus set from 64 to 44 stimuli and the stimulus correlation to r = -0.15 , $\text{BF}_{01} = 1.71$.

In Study 1, the posterior mean of the correlation between the stimulus originality and stimulus utility effects on the drift rate was -0.49, 95% CrI [−0.67, -0.29] whereas in Study 2 it was -0.17, 95% CrI [−0.48, 0.14]. This suggests that the correlation between the originality and utility effects on the drift rate was robust in Study 1 but less so in Study 2.

Probit Model Analysis

To examine whether we could reproduce the main findings using a different, less complex method than the DDM, we conducted a Bayesian hierarchical probit model analysis. The probit model assesses the effect of originality and utility on the proportion of creative responses and therefore disregards response time. We regressed the propensity to respond with "creative" onto stimulus originality and stimulus utility and used a non-informative prior for the intercept and the two effects as well as for the variability parameters, $\text{Normal}(0, 0.3)$. For the correlation across random effects, we used an LKJ prior with shape parameter 3. As can be seen in Figure S7, the extent and nature of individual differences in the originality and utility effects on the propensity to respond with "Yes, creative" is very similar to the DDM results. The posterior mean of the correlation across random effects was similar too: it was -0.43, 95% CrI [-0.62, -0.24] in Study 1 and -0.36, 95% CrI [-0.65, -0.06] in Study 2. These results suggest that the results in the DDM are mainly driven by the task decisions and less driven by response times.

Table S6

Posterior mean, standard deviation of the posterior distribution, 95% credible interval and $\hat{R}$ statistic for the fixed effects (population-level) parameters of the hierarchical probit model analysis (Study 1 and Study 2)

	Study 1				Study 2			
	Mean	SD	LB	UB	Mean	SD	LB	UB
μ	0.12	0.06	0.01	0.23	0.22	0.07	0.09	0.36
μ_{OR}	0.63	0.07	0.49	0.77	0.60	0.08	0.45	0.76
μ_{UT}	0.15	0.07	0.01	0.28	0.15	0.08	0.00	0.31

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

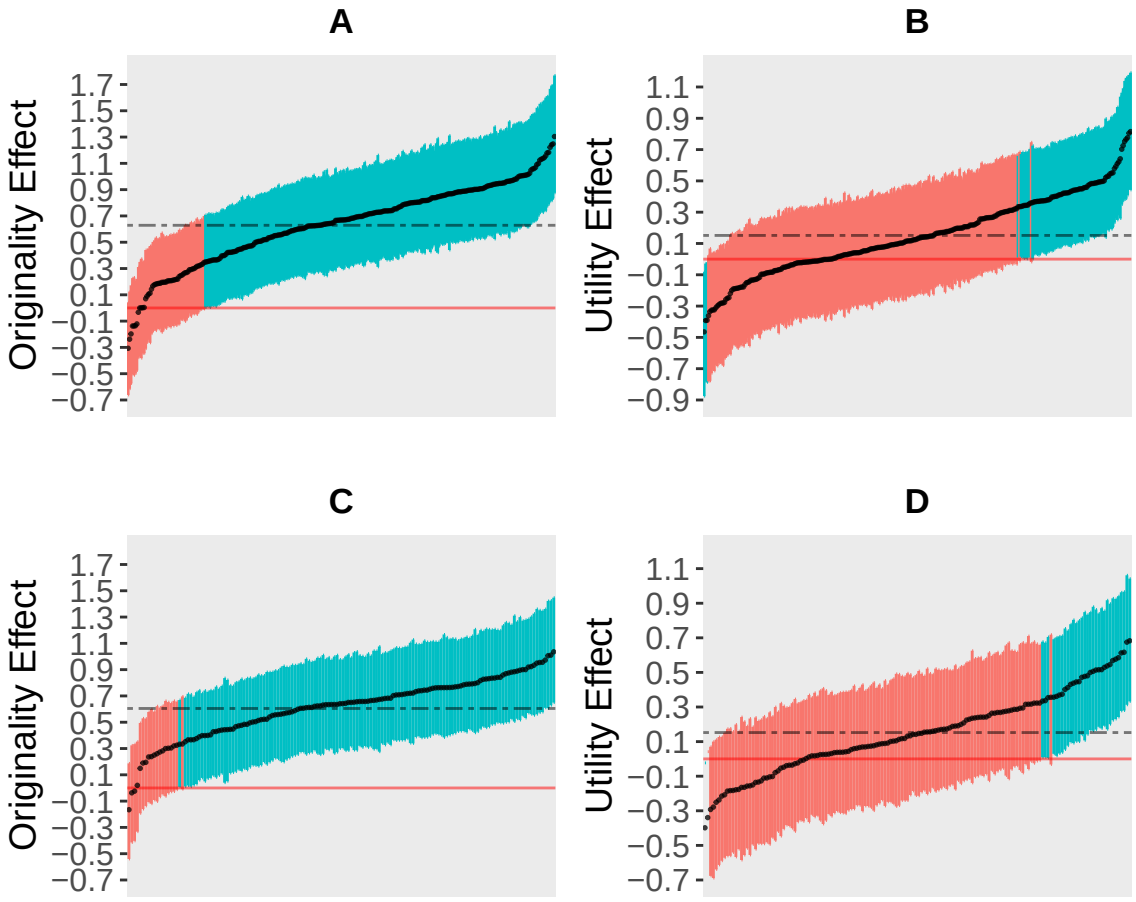

Figure S7. Hierarchical probit model analysis with data from Study 1 (A. and B.) and Study 2 (C. and D.). The plots show the posterior means and the 95 % credible interval (CrI) for each participant in increasing order. The dashed horizontal line denotes the population-level posterior means. CrIs colored in red included zero.

Model comparison using bridge sampling

Given the rather small overall effect of utility on the drift rate, we conducted a Bayesian model comparison to examine whether a simpler model without any utility effects on the drift rate predicted the data better than our specified model with both originality and utility effects on the drift rate. Specifically, we used bridge sampling (Meng & Wong, 1996) and Bayes factors to compare our original model to a model that has no fixed or random

utility effects on the drift rate[1]. Since bridge sampling requires many samples, we re-ran our original model and estimated the less complex model using more than twice as many iterations as in our original analysis (i.e., 30'000 iterations each; post warm-up). To examine the extent of variability in the log marginal likelihoods, we applied the bridge sampler ten times to each fit object. The range of the log marginal likelihoods was -14,629.36 to -14,626.64 for the original model and -14,702.80 to -14,702.15 for the model without the utility effects, which we deemed acceptable. The corresponding Bayes factor range was 1.2e+33 to 4.1e+31 in favor of the original model, overall suggesting overwhelming evidence for our original model compared to a model without utility effects.

References

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

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

Singmann, H. (2018, 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/

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