It is not always better to have more ideas: Serial order and the trade-off between fluency and elaboration in divergent thinking tasks.

Corentin Gonthier, Maud Besançon

It is not always better to have more ideas:

Serial order and the trade-off between fluency and elaboration in divergent thinking tasks

Corentin Gonthier

Université Rennes 2

Maud Besançon

Université Rennes 2

Corentin Gonthier, University of Rennes, LP3C EA 1285, 35000 Rennes, France

Maud Besançon, University of Rennes, LP3C EA 1285, 35000 Rennes, France

Correspondence concerning this article should be addressed to Corentin Gonthier, Laboratoire LP3C, Campus Villejean, Place du Recteur Henri Le Moal, CS 24307, 35043 Rennes Cedex, France. E-mail: corentin.gonthier@univ-rennes2.fr

Abstract

Divergent thinking tasks, which require participants to generate as many creative ideas as possible, elicit a serial order effect: ideas generated later tend to be more original. This suggests that generating more ideas is beneficial. However, past research regarding the serial order effect has largely overlooked the interplay between serial order and fluency: is it always true that more ideas means higher originality? In this study, 595 participants completed four divergent thinking tasks; originality and degree of elaboration were scored for each idea, and multilevel analyses were used to model both originality and elaboration as a function of serial order and total fluency. Later ideas were found to be more original, replicating the serial order effect, but there was an antagonistic effect of sequence length: the ideas of participants with lower fluency tended to be both more original and more elaborate, regardless of serial position. In sum, generating more ideas actually came with lower originality for each idea, despite a serial order effect. These results highlight the role of time and effort for elaboration of an original idea, and also lead to recommending alternate scoring methods in divergent thinking tasks, such as "best-two ideas" or "count of good ideas".

Keywords

Divergent thinking; Originality; Serial order effect; Intraindividual variability; Generalized Additive Mixed Models (GAMM)

Research on creative potential has largely relied on divergent thinking tasks, which require participants to generate as many creative ideas as possible (Guilford, 1950; Runco, 1993), often in a limited timeframe (Kim, 2006). While performance on a divergent thinking task is not isomorphic with "creativity" (see Runco, 2008), divergent thinking is usually considered a good predictor of creative achievement (Kim, 2011). This makes divergent thinking tasks a good framework to explore general mechanisms of creativity. One of the most central questions with divergent thinking tasks, and one that is relevant to creativity in general, is that of the interplay between quantity and quality of ideas. Do individuals who have more ideas also have better ideas?

Relation between Fluency and Originality based on Total Scores and Serial Order

Answering the question of the relation between quantity of ideas (fluency) and quality of ideas (originality) has proven surprisingly difficult. Many studies have examined the relation between fluency and originality aggregated across all ideas. The results often show a positive correlation between fluency and originality (meta-analytic $r = .22$ for average originality in Nijstad et al., 2010), but this is not always the case (e.g. Briggs & Reinig, 2007; Christensen et al., 1957; Nusbaum et al., 2014). Unfortunately, total originality scores are heavily confounded with the total number of ideas in the first place (Reinig et al., 2007; Forthmann et al., 2020; Runco et al., 2008; Silvia et al., 2008). For example, the sum of originality ratings will always be artificially inflated for individuals who generated many ideas; originality averaged across all ideas can be negatively biased when individuals produce a few excellent ideas among many bad ones (Reinig et al., 2007); even a method such as snapshot scoring (Silvia et al., 2008), which rates the global quality of a set of ideas, may be biased by differences of fluency due to differences in cognitive load for raters (Forthmann, Holling, Zandi, et al., 2017). In other words, quantity biases the analysis of quality, which

makes it difficult to determine whether individuals who produce more ideas actually produce better ideas.

Other studies have considered quality at the level of each separate idea, which has led to discovery of the serial order effect: ideas generated later in a sequence tend to be more original, more unique, of higher creative quality. This serial order effect in divergent thinking tasks has been labeled "one of the oldest and most robust findings in modern creativity work" (Beaty & Silvia, 2012), and has been evidenced in many studies (Beaty & Silvia, 2012; Cheng et al., 2016; Christensen et al., 1957; Hass, 2017; Heinonen et al., 2016; Milgram & Rabkin, 1980; Mouchiroud & Lubart, 2003; Parnes, 1961; Phillips & Torrance, 1977; Runco, 1986; Silvia et al., 2017; Wang et al., 2017; Ward, 1969). Prior findings with the serial order effect thus support the notion that it is unequivocally better to always generate as many ideas as possible (for a discussion, see Briggs & Reinig, 2010; Reinig & Briggs, 2013), and that scoring fluency can obviate the need for scoring originality, given that the later ideas will be better than the earlier ones.

However, interpretation of the serial order effect in relation with fluency is not quite as straightforward as could be expected. The serial order does show that originality tends to increase *within a series of ideas*, but it does not directly tell us about the interplay between this serial order effect and the *length of the series*. This begs the question: if later ideas tend to be more original than earlier ones, is it true both for individuals who produce very short and very long sequences? In other words, is the final idea in a 2-ideas-long sequence as original as the final idea in a 20-long sequence – or more, or less? Shorter sequences could be better overall, even if originality tends to increase within a sequence. Again, answering this question is not straightforward, because studies on the serial order effect have often used one of three analytic methods that fail to account for total number of ideas (for a discussion, see Beaty & Silvia, 2012; see also Wang et al., 2017):

1) Comparing ideas in terms of their absolute position in the sequence (comparing ideas generated first, second, and so on; e.g. Christensen et al., 1957; Heinonen et al., 2016), and testing the effect of ordinal position on originality using a form of within-subjects ANOVA. This creates a selection bias: only subjects who have generated ideas in all serial positions retained for analysis can be included (Beaty & Silvia, 2012; Christensen et al., 1957).

2) Dividing the task into several time bins and comparing ideas generated in the various bins (e.g. Cheng et al., 2016; Ward, 1969), which raises a similar issue: subjects who did not generate an idea in any one time bin must be excluded from the analysis (Beaty & Silvia, 2012; Ward, 1969).

3) Comparing ideas in terms of their relative position: for example, comparing ideas generated in the first half versus the second half of the sequence (e.g. Milgram & Rabkin, 1980; Mouchiroud & Lubart, 2003; Parnes, 1961; Phillips & Torrance, 1977; Runco, 1986; Wang et al., 2017). This confounds sequences of ideas with very different lengths ("the second half of the sequence" has a different meaning depending on whether the subject had 2 or 20 ideas).

Another analytic strategy would be possible. Multilevel models (mixed models; e.g. Hox, 2010; Goldstein, 2011) do not suffer from these issues: average originality for the *n*-th idea can be estimated based on the data of all subjects who generated at least *n* ideas, without excluding any subject, and ordinal position can be treated as a numeric predictor without the need for recoding as a categorical variable such as "first half" and "second half" (see Beaty & Silvia, 2012). However, these models have only recently made their way in studies of creativity (Acar & Runco, 2017; Acar, Alabassi, et al., 2019; Acar, Runco, & Ogurlu, 2019; Beaty & Silvia, 2012; Hass, 2017; Kleinkorres et al., 2021; Silvia et al., 2017), and none of

these studies directly tested how serial order combines with fluency, which leaves the question open.

Contrasting Theories Regarding Quantity and Quality in Divergent Thinking

The question of the interplay between quality and quantity has important implications for theories of creativity, given that various theories have made opposite predictions. A first set of theories, although differing wildly in terms of the mechanisms they posit for creativity, have in common the prediction that a higher quantity of ideas should always come with equal or higher quality: in other words, that it is always better to produce more ideas. Another line of thought stresses the role of effort and elaboration in producing good ideas (in contrast with the relatively "passive" view promoted by other theories), and leads to the prediction that there should be a trade-off between quantity and quality and that it could actually be advantageous to producing less, but better ideas.

The first set of theories prominently features the equal-odds rule (e.g. Simonton, 1987, 1997; for another example, see Jung et al., 2015), which proposes that the ratio of good ideas to total number of ideas is constant. It also implies that the total number of good ideas produced by an individual is directly proportional to their total number of ideas (although there should be no increase in average originality as the total number of ideas increases). The quantity-quality conjecture of Osborn (1963; see Reinig & Briggs, 2008, 2013) similarly proposes that the number of good ideas increases with the total number of ideas, and also adds that the average quality of ideas should increase with the total number of ideas.

The list also includes the historically dominant explanation for the serial order effect, the activation spreading account (based on Mednick, 1962). Its rationale is that generating ideas in a divergent thinking task requires activation to be spread in a semantic network centered on the task cue. Semantic nodes closer to the cue would be activated first, eliciting

responses that are highly accessible but not very original; activation would require more time to reach more distant nodes with lower accessibility and higher originality (see Acar & Runco, 2014; Benedek et al., 2012; Milgram & Rabkin, 1980; Wang et al., 2017; Ward, 1969). This account implies that average originality should increase with fluency: the originality of the final ideas in a sequence should be higher when the sequence is longer (as long as activation does not reach nodes that are irrelevant to the task; Reinig & Briggs, 2008).

Lastly, the dual pathway theory (Baas et al., 2013; Nijstad et al., 2010) proposes that high originality can be achieved through flexibility (exploring a large number of idea categories) or persistence (exploring in-depth a few idea categories, which is generally translated as generating many ideas within a few categories; Baas et al., 2013). The dual pathway theory also implies that the number of good ideas should increase with total fluency, as more idea categories are explored or the same category is explored in greater depth.

By contrast, a few theories of creativity predict that quality does not necessarily scale with quantity. Bounded ideation theory (Briggs & Reinig, 2010; Reinig & Briggs, 2008, 2013) proposes that there are limits to the extent to which originality can increase along with total number ideas: originality may well decrease at the end of a sequence of ideas if individuals reach the limit – for example – of their ability, the space of possible solutions to the task, or their willingness to expend effort in the task. Bounded ideation theory does not predict that average originality will *necessarily* be lower for the final ideas in a long series of ideas than in a shorter series, but this is a possible consequence of this view.

Other authors have more explicitly proposed that there can be a trade-off between fluency and originality. This is the case with Guilford (1968; see Forthmann et al., 2020), who argued that individuals who spend their time on producing many (low-quality) responses cannot produce as many good responses. This prediction has been supported by a number of studies finding that subjects instructed to produce more ideas tended to produce ideas of

lower quality (e.g. Christensen et al., 1957; Nusbaum et al., 2014). It is also especially compatible with the finding that quality is correlated with both the time spent on an idea (Acar, Alabassi, et al., 2019; Hass, 2017; Silvia & Beaty, 2012) and its degree of elaboration (Forthmann, Holling, Çelik, et al., 2017; Beaty & Johnson, 2021).

The possibility that original ideas require elaboration is sometimes termed the controlled attention theory of creativity (Beaty et al., 2014). This view stresses the role of top-down control and cognitive abilities such as fluid intelligence (Beaty & Silvia, 2012; Hass, 2017), inhibition (Cheng et al., 2016) and shifting (Wang et al., 2017), and highlights the role of neural activity in "executive" cortical regions in producing a serial order effect (Heinonen et al., 2016; Wang et al., 2017). As stated by Barbot (2018): "creative ideation is an *effortful* process […] it takes more 'effort-time' to come up with an uncommon idea (i.e., involves more exploration/thinking time) than a common one". This theory provides an account of the serial order effect that competes with the activation spreading theory (Beaty et al., 2014): for example, one study indicated that early productions in a divergent thinking task were more likely to be based on a strategy of retrieving ideas in memory, whereas later and more original productions were based on strategies involving more cognitive elaboration (Gilhooly et al., 2007). In this view, it is the time spent on elaborating an idea that matters most, which means generating more ideas can be detrimental if it means these ideas are less elaborate. Note that this can also be reconciled with the dual-pathway theory of creativity (Nijstad et al., 2010), if persistence in a category is considered in terms of time and degree of elaboration rather than the absolute number of ideas generated in this category.

Rationale for the Present Study

The purpose of this study was to leverage multilevel models to investigate the question of the interplay between quality and quantity of ideas, by testing how originality varied as a function of the total number of ideas produced by the participant, and as a

function of serial order – that is, at the level of each separate idea in a sequence. We expected the results to answer the question of whether it is always better to have more ideas, as suggested by much of the literature, and therefore to provide insight into competing theories of creativity.

Indeed, the theories outlined in the previous section lead to very different predictions regarding the combination of fluency and serial order. Briefly, the equal-odds rule leads to the prediction that average originality should be stable regardless of fluency and serial position. Both the quantity-quality conjecture and the activation spreading account predict that originality should be highest for the final ideas of participants who generate a large number of ideas. By contrast, the controlled attention theory of creativity makes the very different prediction of a trade-off between quantity and quality: there should be a serial order effect, but originality and elaboration should also be substantially higher for short sequences than for long sequences, which means originality should be highest at the end of short sequences. In other words, testing whether there is a trade-off between quantity and quality in association with the serial order effect (whether originality is better for short sequences or at the end of long sequences) should provide a window into the relative role of effortful processes.

Bounded ideation theory predicts that average originality could possibly decrease at the end of long sequences, which fits rather well with the controlled attention view of creativity, but it does not make specific predictions in terms of fluency. The dual pathway theory is also a bit of a mixed case: originality should increase as a function of fluency and serial order, unless "persistence in a category" is taken in terms of invested effort rather than the absolute number of ideas, in which case there could also be a benefit to generating short sequences.

To investigate the interplay between fluency and serial position, we had a large sample of participants complete two divergent thinking tasks (adapted from Lubart et al., 2011, and requiring creativity with verbal and scientific materials), with two sessions for each task. We assessed the originality of each creative production, along with their degree of elaboration as indexed by their length (number of words). We then tested how originality varied as a function of both serial order, total number of ideas, and the interaction of the two. Critically, we also performed the same analyses with elaboration, as a window into whether fluency also affected elaboration, as predicted by the controlled attention theory.

The current study used a particular form of multilevel models: generalized additive mixed models (GAMM models; see Wood, 2017). This extension of multilevel models makes no particular assumption on the shape of the relationship between the dependent variable and its predictors, and thus allows for the estimation of non-linear relationships of any shape. This approach is particularly suited to the study of intra-individual variability when there is no *a priori* hypothesis on the shape this variability can take (for an example, see Gonthier & Roulin, 2019). It is also particulary well-suited to the creation of heat-maps reflecting scores at the combination of two variables (for an example, see Figure 1C), and therefore particularly well-suited to examining how the effects of fluency and serial order add up. This analytic strategy influenced the design of the study: GAMM analyses require a large sample size, and they can be vulnerable to overfitting, which drove us to analyze the relation between serial order and total sequence length in a large dataset (collected across two testing sessions), and across two divergent thinking tasks as a form of cross-validation.

Method

Participants

Data were collected from 595 students enrolled in the same computer science school at the undergraduate level (96 females and 499 males; mean age = 23.72 years, *SD* = 4.24). Participants were recruited over two school years. All participants were native French speakers, and none had completed any of the experimental tasks before. All participants provided written informed consent prior to the experiment.

Materials

Participants were invited to complete two different divergent thinking tasks, adapted from the Evaluation of Potential of Creativity (EPoC: Lubart et al., 2011) task battery. Participants were asked to complete each task twice, with slightly different materials for the first and second session. The first task assessed creativity in the verbal domain, by asking participants to imagine endings (session 1) or beginnings (session 2) for a story. The second task assessed creativity in the scientific domain, by asking participants to imagine explanations for a phenomenon observed in the field of human sciences (session 1) or physics (session 2). Each task had two forms (form A and form B), identical except for the specific prompt about what exactly the participant was required to imagine (e.g. the specific incomplete story for the verbal task). Each participant was randomly assigned to complete one of the two forms.

The instructions were identical for all tasks: before reading the prompt for a given task, participants were told that: "You should propose multiple ideas, and you should propose ideas that are original and different from what others may propose. You have ten minutes to propose as many ideas as possible". The tasks were computerized, so that participants typed their responses in a blank response field. They could always see all their prior responses throughout a task.

Procedure

Testing took place online and could be completed at any time during a one-month window. Participants were invited to complete the two testing sessions through e-mails sent by their school. After a personality questionnaire not reported here, participants completed the divergent thinking tasks for the verbal, then scientific domains. The whole experimental session lasted approximately 40 min. The second session was identical but included only the creativity tasks.

Participation was completely voluntary, and not all students completed all tasks. Apart from the dataset of 595 students analyzed in this study, another 182 students clicked on the link for the study, but provided no scorable answers or discontinued testing without completing the first task. Among the 595 students who contributed data, 142 completed only the first task of the first session; 95 completed two tasks; 92 completed three tasks; and 266 completed all four tasks. The data appeared to be missing relatively at random (there were no relations with age, gender, or performance, all rs < .10).

Scoring

The responses from all participants in the creativity tasks were retrieved and segmented into separate productions (for example, "maybe old people are tired, or maybe they have lived long enough to understand that they have time to do things" was counted as two ideas).

The originality of each production was scored on a three-point scale (low, medium, or high originality) by raters. There were three different raters: the second author (a co-creator of the EPoC) and two graduate students (blind to the study hypotheses regarding the relation between fluency, originality and elaboration). The two graduate students who served as raters were purposely trained to score creativity tasks: they completed a 2-hour training session including a series of exercises in scoring divergent thinking tasks, where they received

detailed feedback about their scoring, and where it was ensured that they had at least 75% absolute agreement with the average of all raters in the EPoC norming data (see Lubart et al., 2011). One grad student scored creativity in the verbal domain, the other scored creativity in the scientific domain, and the second author scored all tasks. This means each production was scored by two raters, whose estimates were averaged.

Inter-rater agreement in the current study was ensured by having the three raters score originality for the same set of 300 productions, randomly drawn from the whole dataset. The three judges then discussed their scoring on this sub-sample until consensus was achieved, before proceeding to score their assigned portion of the dataset. Reliability was then estimated by computing an intra-class correlation coefficient between the two raters, on the whole sample (ICC 3,1: two-way mixed, average measures, absolute agreement; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Reliability was ICC = .97 for both tasks, representing excellent inter-rater agreement (Cicchetti, 1994).

Elaboration was scored by counting the number of words in each production (for a discussion, see Dumas et al., 2021). This was done using R (R Core Team, 2021). Text strings were first pre-processed by removing special characters such as punctuation, numbers or dual spaces; we also screened them manually for artifacts. Flexibility was also scored by a single rater, who scored the category corresponding to each idea; these data are reported as descriptive statistics (total number of categories generated by the participant) but they were not analyzed for the present study.

Data Analysis

The data were analyzed using GAMM (see Wood, 2017). Statistical inference was performed based on approximate *p*-values. For each predictor, we report the *F* decision statistic, the corresponding *p*-value, and the effective degrees of freedom (*edf*: effective

degrees of freedom equal to 1 reflect a linear relationship between predictor and dependent variable, values greater than 1 reflect a more complex trajectory).

GAMM analyses were performed using the *mgcv* package (Wood, 2017; version 1.8-28) for *R* (R Core Team; version 3.6.1), along with the *itsadug* package for figures (van Rij et al., 2020). Data analysis was performed at the level of ideas in the task. The originality rating of an idea was modeled assuming an ordered categorical distribution, and the number of words was modeled assuming a gaussian distribution. The numbers of words of each idea was log-transformed (natural logarithm of the number of words plus one) prior to statistical inference to account for positive skewness (skewness was 2.74 before and 0.11 after log-transformation; kurtosis was 11.16 before and -0.10 after). Subject-level random effects were modeled as random intercepts, plus random slopes for the effect of idea serial position. All analyses tested the main effect of ordinal position of the idea, the main effect of the total number of ideas generated by the participant, and the interaction between the two. The type of task (verbal versus scientific), session (session 1 versus session 2) and form of the test (form A versus form B) were also added as nuisance covariates[1].

Models were fit using restricted maximum likelihood. Smooths were modeled with the default classes – thin plate regression splines for simple smooths and cubic regression splines for interactions. Basis dimension was fixed at $k = 6$, which was sufficient for all analyses (as determined based on inspection of k-indices and refitting the models when the *edf* were close to the maximum allowed value; Wood, 2017). The results were also visually inspected to ensure the absence of overfitting.

[1] The model formulas were of the form Originality ~ Task*Session*Form + s(IdeaSerialPosition, k=6) + s(Fluency, k=6) + ti(IdeaSerialPosition, Fluency, k=6) + s(SubjectID, bs="re") + s(SubjectID, IdeaSerialPosition, bs="re"). This syntax tests nonlinear effects of serial position, fluency, and the interaction between the two, while allowing for average differences between tasks (verbal vs. scientific), sessions (session 1 vs. session 2), and forms (form A vs. form B), also including subject-level random intercepts and random slopes for serial position.

Results

Preliminary Analyses

The sample included $N = 595$ unique participants completing one or multiple creativity tasks, which together amounted to 1672 task sessions. Among these 1672 task sessions, there were a total of 481 sessions where participants only proposed a single idea, which were removed from the dataset. There were also 19 sessions where participants proposed more than 13 ideas; these sessions were kept, but only the first 13 ideas were analyzed[2]. This resulted in a dataset of 1191 task sessions for a total of 5559 ideas. Descriptive statistics are displayed in Table 1.

Bivariate correlations between all measures aggregated at the task level are also displayed in Table 2 (these correlations were computed based on all task sessions). As discussed above, analyses on total scores cannot replace analyses performed at the level of each idea (as reported in the next session). They do however provide a first hint of a tradeoff between quantity and quality in the current dataset: the total number of ideas was negatively correlated with both average originality and average elaboration. This held true even when analyzing the tasks separately: originality was negatively correlated with fluency in all four tasks (all rs < -.10, all ps < .05), as was elaboration (all rs < -.30, all ps < .05).

[2] There were $n = 33$ sessions with 13 ideas, which seemed to be the minimum reasonable value for estimation. There were only $n = 19$ sessions with 14 ideas or more.

Table 1

Descriptive statistics for all measures

Creativity task	*N*	Total number of ideas			Total number of categories			Originality per idea			Number of words per idea		
		M	*SD*	range	*M*	*SD*	range	*M*	*SD*	range	*M*	*SD*	range
Verbal S1	355	4.12	2.56	2-13	3.20	1.32	1-7	1.62	0.77	1-3	35.85	33.68	2-294
Verbal S2	245	3.12	1.68	2-13	2.24	0.84	1-4	1.63	0.76	1-3	44.05	37.42	1-254
Scientific S1	320	6.02	2.99	2-13	4.49	1.70	1-9	1.65	0.65	1-3	19.81	15.60	1-155
Scientific S2	271	5.19	2.70	2-13	3.70	1.61	1-8	1.47	0.64	1-3	19.12	16.05	1-154

Note. S1 = session 1 ; S2 = session 2 ; *N* = total number of participants who completed this task ; *M* = mean ; *SD* = standard deviation.

Table 2

Bivariate correlations between all measures at the task level

Measure	Total number of ideas	Total number of categories	Average originality	Average number of words per idea
Total number of ideas	-	.75	-.15	-.53
Total number of categories	.75	-	-.08	-.40
Average originality per idea	-.15	-.08	-	.42
Average number of words per idea	-.53	-.40	.42	-

Note. $N = 1191$. All correlations are significant at the $p < .01$ level.

Originality as a Function of Serial Order and Fluency

A first series of analyses investigated the serial order effect for originality, by testing whether originality varied as a function of the ordinal position of the idea in the sequence generated by the participant, as a function of the total number ideas generated by the participant (fluency), and the interaction between the two.

The results are displayed in Figure 1. There was a significant effect of idea ordinal position, $F = 20.55$, $edf = 1.00$, $p < .001$: as displayed in Figure 1A, ideas generated later in a sequence tended to be substantially more original than ideas generated earlier, indicating a

serial order effect. However, there was also an antagonistic effect of fluency, $F = 86.75$, $edf = 1.00$, $p < .001$: as displayed in Figure 1B, the ideas of participants who generated few ideas were, on average, much more original than the ideas of participants who generated many ideas. The interaction between ordinal position and total fluency was also significant, but of low magnitude, $F = 2.76$, $edf = 3.34$, $p = .033$. The combined effect of the two variables is represented in Figure 1C. The interaction between ordinal position and fluency reflected a slight nonlinear serial order effect for participants with very high fluency, compatible with a slight decrease in originality at the end of very long sequences, but not critical to interpretation of the data.

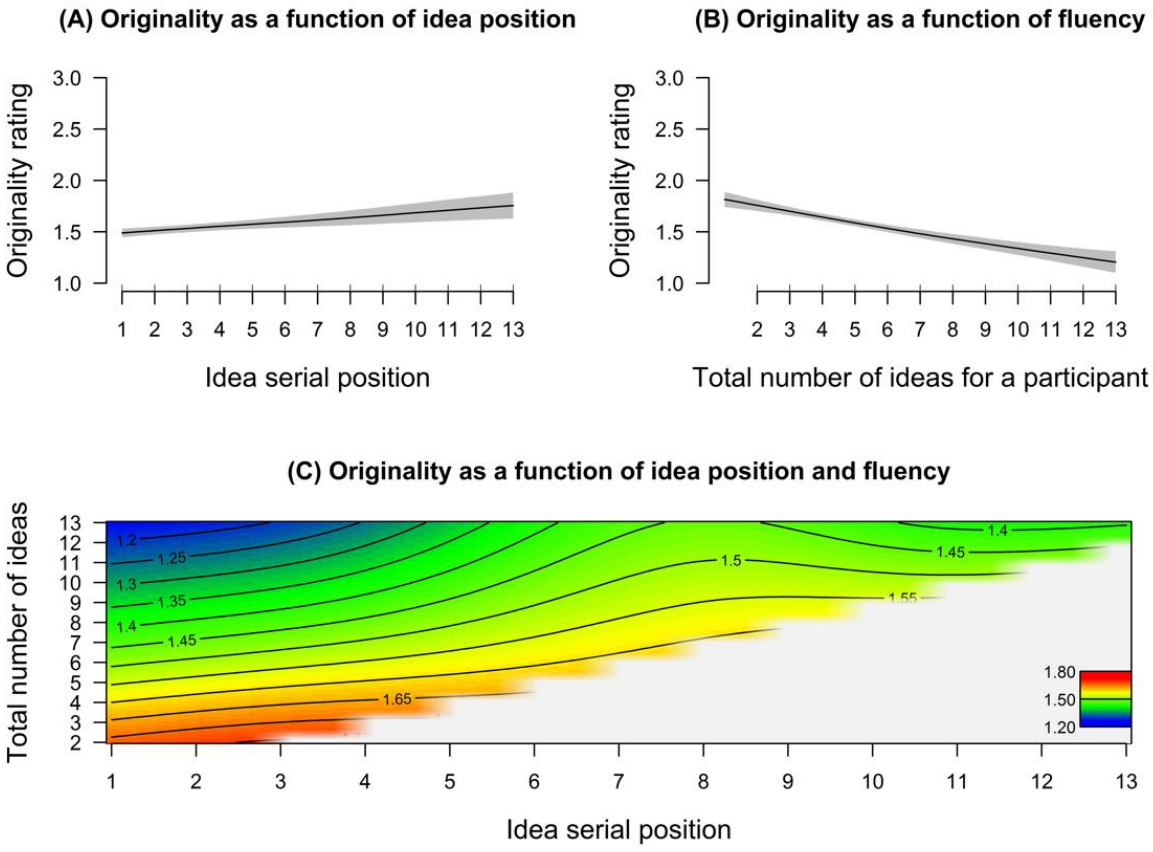

Figure 1. Originality as a function of the serial position of an idea (A), fluency (B), and the combination of the two variables (C). Panel C is easiest to read horizontally – for example, the average originality of the first idea was approximately 1.70 for a participant who generated 2 ideas in total (bottom left), and approximately 1.15 for a participant who generated 12 ideas in total (top left). The greyed-out area does not have data (e.g. it is not meaningful to estimate originality for the 13th idea of a participant who generated two ideas).

As can be seen in Figure 1C, on average, the highest creativity scores were obtained for the ideas of participants who generated very short sequences (bottom left of the figure, red color). In the case of participants who generated long sequences of ideas, the first ideas had very low originality ratings (top left of the figure, blue color); originality did substantially increase throughout the sequence, but the final ideas had only average originality ratings (top right of the figure, green color), never quite reaching the originality of those participants who generated much shorter sequences. In other words, we did clearly replicate the serial order effect, but the increase in originality throughout a sequence was overshadowed by the fact that shorter sequences were overall much more original, indicating a quantity-quality trade-off.

Of secondary interest, a complementary analysis was performed to confirm that this pattern of results was stable across the four tasks that composed the current dataset: creativity in the verbal and scientific domain, each performed across two sessions. This analysis compared models allowing, or not, the effects of serial order and fluency to differ as a function of task and session. The best fit was obtained for a model where the effects of serial order and fluency varied across tasks, but not their interaction (see Table 3; Model 4). This model indicated that the main effects of serial order and fluency were of different magnitude in the different tasks, but this variation explained very little deviance in the model; the interaction between serial order and fluency was not significantly different across tasks.

The corresponding results (with main effects and an interaction between serial order and fluency, and the two main effects allowed to vary across tasks) are displayed in Figure 2. The pattern of results was in fact very similar in all four tasks. There was a significant serial order effect reflecting increasing originality throughout serial positions in all tasks (all ps < .001), except for the first session of the verbal task where the effect was non-significant, $F = 0.32$, $edf = 1.11$, $p = .681$. There was also a significant negative effect of total fluency in

all tasks (all *p*s < .010). In all cases, the highest originality ratings were obtained for participants with very low fluency, whereas the lowest ratings were obtained for the first ideas of participants with high total fluency, again reflecting a quantity-quality trade-off. Despite increasing creativity throughout serial positions, the final ideas of participants with high total fluency never quite reached on average the same originality as the ideas of participants with very low fluency, except in the first session of the verbal task. The interaction between fluency and serial position did not reach significance in any of the four tasks when considered separately, all *p*s > .20.

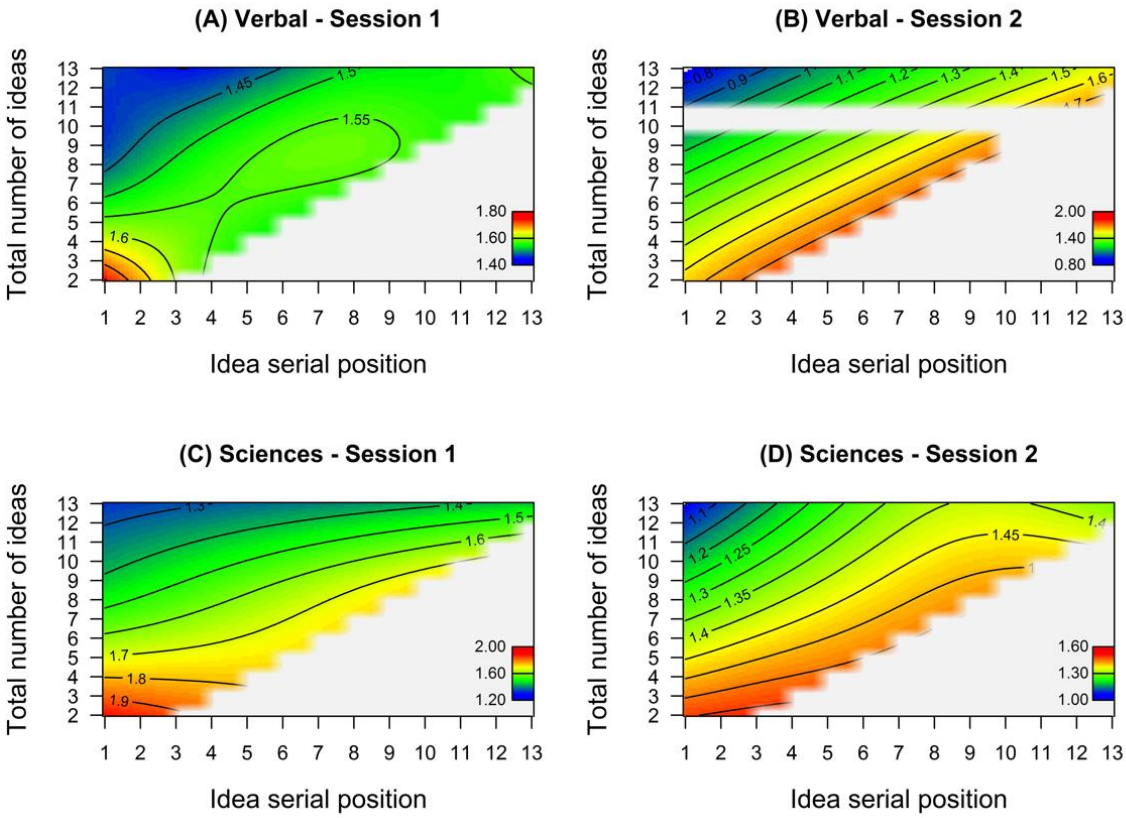

Figure 2. Originality as a function of serial position and fluency, as estimated separately for the four experimental tasks. See Figure 1 for details on how to read these plots.

Table 3

Model comparison testing the moderation by type of task for originality

Model	Moderation by task and session	Model fit (AIC)	Percent explained deviance
Model 1	None	9030	25.0%
Model 2	Serial order	9021	25.3%
Model 3	Fluency	9029	25.1%
Model 4	Serial order + Fluency	9020	25.3%
Model 5	Serial order + Fluency + Interaction	9022	25.4%

Note. AIC = Akaike Information Criterion; a lower value indicates better fit. All models were based on the formula detailed in Footnote 1.

Elaboration as a function of Serial Order and Fluency

The next series of analyses explored the role of elaboration as a possible explanation for the higher originality ratings of participants with low fluency: with the same amount of time to spend on less ideas, these participants may have had more time to elaborate more complex or detailed ideas, possibly eliciting higher originality ratings (as advocated by other studies: Forthmann, Holling, Çelik, et al., 2017; Beaty & Johnson, 2021). The analyses for elaboration were identical to those conducted for originality ratings, except that originality was replaced with the (log-transformed) number of words as a dependent variable.

The results are displayed in Figure 3. There was a significant effect of idea ordinal position, $F = 6.72$, $edf = 3.88$, $p < .001$: as displayed in Figure 3A, the effect had a nonlinear shape wherein ideas generated in the middle of a sequence were on average somewhat more elaborate than ideas generated earlier or later. There was also a very large negative effect of fluency, $F = 57.55$, $edf = 3.53$, $p < .001$: as displayed in Figure 3B, the ideas of participants who generated few ideas were, on average, much more elaborate than the ideas of participants who generated many ideas. The interaction between ordinal position and total fluency was not significant, $F = 1.64$, $edf = 4.40$, $p = .148$, indicating that these effects were additive. The combined effect of the two variables is represented in Figure 1C. As can be seen, the highest elaboration was obtained for the ideas of participants who generated very

short sequences. The ideas proposed by participants who generated long sequences were much less elaborate, and elaboration did not substantially increase with increasing serial position. (To illustrate, the first idea of participants who generated two ideas contained on average 62.66 words, whereas the first idea of participants who generated 13 ideas contained on average 10.12 words.)

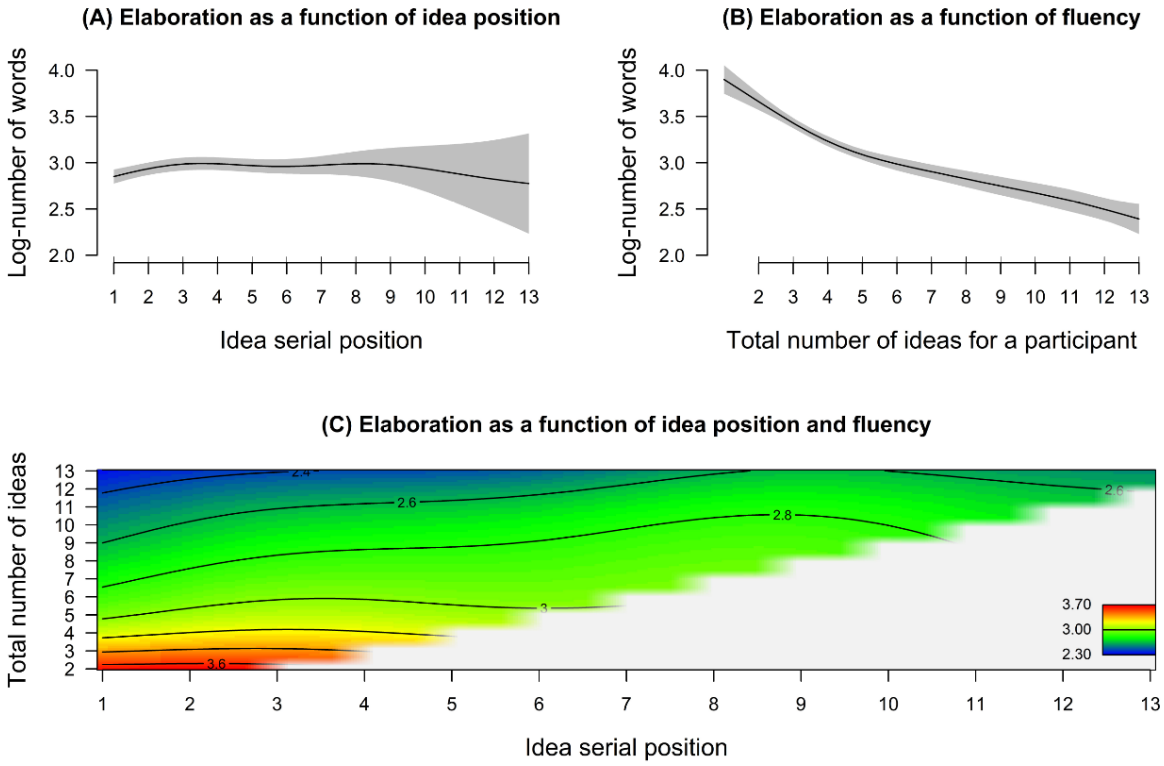

Figure 3. Elaboration, expressed as the logarithm of the average number of words in an idea, as a function of serial position of an idea (A), fluency (B), and the combination of the two variables (C). See Figure 1 for details on how to read these plots.

A complementary analysis was again performed to confirm that this pattern of results was stable across the four tasks. The best fit was obtained for a model where all effects varied across tasks, though this variation across tasks again explained very little deviance (see Table 4; Model 5). The corresponding results are displayed in Figure 2. As was the case for

originality, the pattern of results was very similar in all four tasks: the negative main effect of fluency on elaboration was significant in all tasks (all *p*s < .001), indicating that the ideas of participants with low fluency were always more elaborate (see Figure 2). Of lesser interest, the main effect of serial position was significant or at the trend level in all cases (all *p*s < .09), but never offset the detrimental effect of fluency. The interaction was significant only for the first session of each task (both *p*s < .05).

Table 4

Model comparison testing the moderation by type of task for elaboration

Model	Moderation by task and session	Model fit (AIC)	Percent explained deviance
Model 1	None	9251	67.2%
Model 2	Serial order	9243	67.4%
Model 3	Fluency	9238	67.2%
Model 4	Serial order + Fluency	9227	67.4%
Model 5	Serial order + Fluency + Interaction	9213	67.6%

Note. AIC = Akaike Information Criterion; a lower value indicates better fit. All models were based on the formula detailed in Footnote 1.

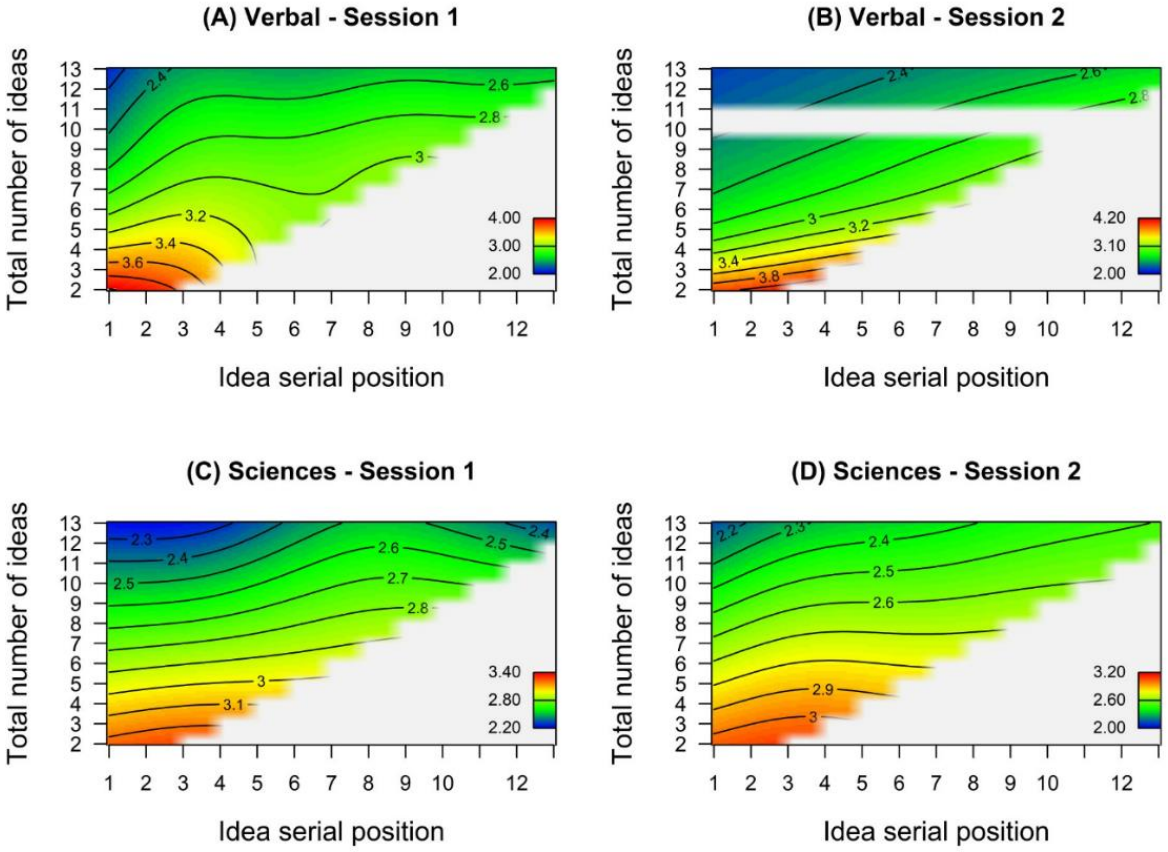

Figure 4. Elaboration as a function of serial position and fluency, as estimated separately for the four experimental tasks. See Figure 1 for details on how to read these plots.

Discussion

This study aimed to examine the interplay between quantity and quality in divergent thinking tasks, by testing originality and elaboration as a function of fluency and serial order. The results revealed two major findings: firstly, we found a serial order effect similar to past literature, but the effect was overshadowed by a strong negative effect of fluency, so that originality was substantially higher for participants who generated fewer ideas. Originality did increase with serial position for participants who generated longer sequences, but it remained generally below the originality ratings of participants with low fluency. Secondly, these results for originality were associated with a similar pattern for elaboration, as reflected in word count: elaboration was substantially higher for participants who generated fewer ideas (for similar results, see Forthmann, Holling, Çelik, et al., 2017; Dumas et al., 2021). Elaboration was somewhat related to serial order, but to a lesser extent than originality, and elaboration for participants with high fluency remained clearly below participants with low fluency for all serial positions. These patterns appeared to be stable across the four tasks investigated here.

In short, there was a substantial trade-off between quantity and quality of ideas: participants who had less ideas proposed ideas that were both more original and more elaborate than participants who had more ideas[3]. This was true despite the existence of a significant serial order effect: the ideas of participants who generated fewer ideas were always equal or better than the final ideas of participants who generated long series of ideas. These results, while replicating the often-studied serial order effect, paint it in a different light and lead to a conclusion that could appear strongly counter-intuitive to researchers used to the

[3] A latent profile analysis performed for exploratory purposes, as suggested by a reviewer, showed that there were no clear-cut profiles of participants in the task (such as two profiles of "many responses with low originality and low elaboration" and "few responses with high originality and high elaboration"). In other words, there was no definite profile of mass-producers (Feist, 1997). The data were better summarized as a continuous distribution of fluency, originality and elaboration, with fluency inversely related to originality and elaboration.

serial order effect - or used to scoring only fluency in divergent thinking tasks. Despite the existence of a serial order effect representing the fact that the final ideas in a sequence do indeed tend to be better than the first ideas, generating fewer but more elaborate ideas is in fact associated with higher originality – at least for the tasks presented here. In other words, these results clearly show that it is not always better to have more ideas.

Theoretical Consequences for Accounts of Divergent Thinking Tasks

Our results are clearly incompatible with the quantity-quality conjecture that average originality should increase with the total number of ideas (Osborn, 1963; see Reinig & Briggs, 2008, 2013), and they also run counter to the equal-odds rule (Simonton, 1997) stating that average originality should not depend on the total number of ideas. This rule was developed based on the analysis of lifetime productions of creators such as composers, but constraints on the creative process mean that it does not necessarily hold in more bounded tasks, such as time-constrained divergent thinking tasks (see Briggs & Reinig, 2010).

By confirming the existence of a trade-off between quantity and quality that is sufficient to overshadow the serial order effect, the pattern of results observed here instead clearly supports those accounts of divergent thinking that stress the role of cognitive abilities and active elaboration over time in producing good ideas, such as the controlled attention theory (Beaty & Silvia, 2012; Beaty et al., 2014; Cheng et al., 2016; Christensen et al., 1957; Gilhooly et al., 2007; Guilford, 1968; Hass, 2017; Heinon et al., 2016; Silvia & Beaty, 2012; Wang et al., 2017). The results for elaboration directly support the view that more elaboration yields better ideas, that the originality results are dominated by the role of elaboration, and that elaboration of an idea requires sufficient time – and thus that less ideas be produced in total. While there was only weak evidence

It is possible to reconcile our findings with the dual pathway theory of creativity, which proposes that high originality can result from persistence (Nijstad et al., 2010).

However, this requires a loose interpretation of "persistence" – in the sense of investing a high "degree of sustained and focused task-directed cognitive effort" (Nijstad et al., 2010, p. 42), which could conceivably be extended to elaboration (see also Taylor et al., 2021). The usual interpretation of persistence in the sense that "ideas within a particular category become more original as time proceeds and more ideas have been generated within that category" (p. 47) cannot explain why originality was higher for ideas produced in very short sequences. Note that an alternative interpretation of our results could have been that participants producing few ideas actually generated as many ideas as the others, but had a higher threshold for including them as responses (i.e. they could have censored most of their ideas, only keeping the best ones), explaining their higher originality; but the fact that their ideas were also much more elaborate makes it clear that in our case effective "persistence" was actually about investing time and effort into an idea rather than producing many ideas and censoring most of these.

On the surface, our results also seem incompatible with the activation spreading account. Based on activation spreading, we would have expected the best ideas to appear at the end of long sequences, reflecting the fact that activation reached very distant semantic nodes – not in the productions of participants who proposed only a couple of ideas. This conclusion converges with prior studies finding that proposing more ideas tends to come with diminishing returns, incompatible with predictions based purely on activation spreading (Beaty & Silvia, 2012; Kleinkorres et al., 2021; Reinig & Briggs, 2008). In fact, it is possible that activation spreading also plays a role and also contributes to the serial order effect observed here: the roles of associative and executive processes can be reconciled in dual-process accounts (Beaty et al., 2014; see also Barr et al., 2014). What our results do show is that elaboration plays a substantial role, given that it would be impossible to explain the pattern of results purely through activation spreading; and given the fact that fluency had

more effect than serial order, we would conjecture that the role of elaboration is greater than that of activation spreading.

Practical Consequences for Scoring of Creativity Tasks

Our results unequivocally speak against the practice of scoring only fluency in creativity tasks (e.g. Diehl & Strobe, 1987; Hargreaves & Bolton, 1972). Indeed, generating many ideas is no guarantee that any of them will be good. This is easily illustrated in the current dataset: the participant with the highest fluency proposed 23 ideas in the same task, but ideas 14 to 23 were on average three words long, and they all received the lowest originality score. In other words, scoring only fluency may be much easier than having judges assess the originality of each discrete idea (Amabile, 1982), but it can mask the fact that the ideas are of very poor quality. Our results help piece together why fluency can be a poor measure of creativity, but the idea itself is not novel: for example, Simpson (1922) argued that the number of productions "is no index of a person's creative imagination" (for similar arguments, see Baer, 2011; Zeng et al., 2011).

A number of classic tests of divergent thinking (such as the Torrance Tests of Creative Thinking; see e.g. Zeng et al., 2011) solved this problem by scoring all of fluency, originality, elaboration, and flexibility, and encouraged testers to interpret the combination of these scores. While this is certainly a possibility, this solution is highly resource-intensive, which can explain why some researchers default to scoring only fluency. Some scores may also be difficult to determine (such as elaboration; Cramond et al., 2005), and the combination of originality, fluency, elaboration and flexibility is difficult to interpret, even when these dimensions are integrated in composite scores (which means a high score can be obtained with very different response patterns). Furthermore, rating originality or flexibility in addition to fluency does not solve the fact that aggregate scores are heavily confounded with differences of fluency (Forthmann, Holling, Zandi, et al., 2017; Forthmann et al., 2020;

Reinig et al., 2007). This is true even when these indices are combined: this gives a heavy weight to fluency since it tends to confound other measures (e.g. Forthmann et al., 2020).

Several authors have proposed alternative coding schemes that are much less vulnerable to the confounding between fluency and originality (for a discussion, see Forthmann et al., 2020). A promising solution is "Top 2 scoring" (Silvia et al., 2008), where participants are asked to choose their best two ideas, and only these two ideas are scored. This method is entirely independent of fluency, and has the added benefit of decreasing the number of ideas to be scored for originality by the tester, although it makes scores less tractable by confounding the quality of ideas with the ability to identify one's good ideas (see also Runco, 2008). Another example is "count-of-good-ideas" scoring (Briggs & Reinig, 2010; Reinig et al., 2007), which simply counts how many good ideas (ideas above a certain threshold of quality) were generated by the participant. This method is not entirely unconfounded by fluency, but it does take originality into account and avoids overestimating the scores of participants who generated a large number of low-quality ideas, as would be the case with sum-of-originality scoring (or underestimating their scores, as would be the case with average-originality scoring). This could make it an appropriate approach for applied settings where the total number of good productions matters more than the relative performance of a given individual. Either method reduces the interpretation problem posed by participants with high fluency and low elaboration.

Limitations and Directions for Future Studies

While we took steps to ensure the stability of our findings across four different tasks, our results leave open a major question about generalizability. The four tasks had the same structure, and they were performed by the same sample of undergraduate students in computer science, under the same conditions of online testing. This leaves open the possibility that the results were driven by particular aspects of the methods. Obtaining

different results under different testing conditions would not directly question our conclusion that it is not *always* better to have more ideas, but it would be interesting to determine how common this pattern can be in other tasks and situations. We believe two particular aspects of the procedure require further discussion.

A first point of discussion concerns the way participants understood the requirements of the task. The instructions used in this study encouraged participants to produce "as many original ideas as possible", which stresses both fluency and originality. This is an example of "hybrid instructions" (Reiter-Palmon et al., 2019), which may have been perceived differently by different participants, leading them to stress either quantity or quality as a function of their response style (e.g. Baer, 2011). In this sense, the instructions may have encouraged inter-individual variability in fluency and elaboration/originality, in essence creating a sort of speed-accuracy trade-off. In parallel, the absence of direct human supervision may have led participants to place less emphasis on the quality than the quantity of their responses, possibly yielding a larger number of responses with low elaboration. The compromise between fluency and elaboration/originality observed here should exist in all situations, but it could be more difficult to observe – or have less influence on the results – in a task explicitly stressing quality over quantity and performed under direct human supervision.

A second point of discussion concerns the structure of the task itself. Our divergent thinking tasks required production of original ideas relating to stories and hypotheses about the world; this is rather different from the unusual uses task which has been used in many of the prior studies about divergent thinking (e.g. Beaty & Silvia, 2012). Moreover, creative performance, and originality in particular, tends to be largely task-specific and to correlate poorly even across versions of similar divergent thinking tasks (Barbot et al., 2019). On the other hand, there is little reason to believe that a compromise between fluency and elaboration should be dependent on the type of materials to be produced: the effect seems to

be about how participants invest limited time in a task, more than interaction with a specific topic. We therefore expect that the same effect would be observed in other divergent thinking tasks.

Much more important than the content of the tests should be the timing of the task. In this case, the presence of a 10-minutes time constraint per task may have influenced the sequences of ideas in particular ways: a shorter time constraint could have constrained fluency to the point of eliminating the trade-off with elaboration. It is however doubtful whether a longer time constraint, or no contraint at all, would have made the trade-off disappear: it is possible that participants self-impose a time limit on a task, as driven by cognitive fatigue or motivation to perform well (see Briggs & Reinig, 2010), leading to the same compromise between quality and quantity even in the absence of an explicit time limit. For example, the length of the whole procedure in the current experiment may have played a role in discouraging participants.

In fact, there is no way with the present dataset to tell whether the limit on elaboration for participants who produced more but less elaborate ideas was actually the time limit on the task. The compromise that some participants appear to have made by trading high fluency with low elaboration may have been driven by limited motivation rather than insufficient time to produce better ideas. It is also possible that some participants lacked the ability to elaborate, due to lower cognitive skills (see Beaty & Silvia, 2012; Briggs & Reinig, 2010).

This leads us to what we believe to be the most promising direction for future studies: exploring in greater detail the role of time, and more specifically latency (time spent on creating each idea). Research about divergent thinking has often measured the originality, flexibility and elaboration of each idea, but it has rarely been interested in latency, and especially the way latency can vary throughout a series of responses. Examples include the works of Hass (2017), showing temporal and semantic clustering of responses, and Acar and

colleagues (Acar & Runco, 2017; Acar, Runco, & Ogurlu, 2019), showing that changes of response categories are associated with longer latencies, especially towards the end of a sequence. A methodological benefit of studying latency in addition to serial order is that assessing creativity as a function of time intervals is that it can partly mitigate the problem of low precision of estimates for higher serial positions (e.g. Beaty & Silvia, 2012; Kleinkorres et al., 2021), due to the low number of participants generating data at these positions (see Footnote 2 and Figure 3A for examples). Theoretically, latency may also be more relevant than fluency, for example to assess the degree of persistence in exploring a particular idea category (Nijstad et al., 2010).

Given our data, it seems likely that participants with low fluency spent a lot of time on just a few ideas, whereas participants with high fluency spread their time evenly across a large number of less-elaborate ideas; unfortunately, we did not log response times and this could not be verified. It would also be interesting to determine whether the classic serial order effect for originality is accompanied by changes in response times (when originality increases with serial position, is it also the case that time spent on each idea systematically increases, or decreases, throughout the sequence? how does this relate to elaboration?). Generally speaking, further exploring the role of ideation time in divergent thinking tasks would allow for a more precise understanding of how and why idea quality changes across serial positions, by offering another window into the role of fatigue, elaboration, time constraints, and the limits of the solution space (Briggs & Reinig, 2010).

Acknowledgments

We thank the school who volunteered to participate in the data collection, and anonymous reviewers for their very detailed and insightful comments on the manuscript.

References

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal, 26*(2), 229–238. doi:10.1080/10400419.2014.901095

Acar, S., & Runco, M. A. (2017). Latency predicts category switch in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 43–51. doi:10.1037/aca0000091

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

Acar, S., Runco, M. A., & Ogurlu, U. (2019). The moderating influence of idea sequence: A re-analysis of the relationship between category switch and latency. *Personality and Individual Differences*, *142*, 214–217. doi:10.1016/j.paid.2018.06.013

Amabile, T.M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*, 997-1013.

Baas, M., Roskes, M., Sligte, D., Nijstad, B. A., & De Dreu, C. K. W. (2013). Personality and creativity: The dual pathway to creativity model and a research agenda. *Social and Personality Psychology Compass*, *7*, 732–748. doi:10.1111/spc3.12062

Baer, J. (2011). How divergent thinking tests mislead us: Are the Torrance tests still relevant in the 21st century? The Division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(4), 309 –313. doi:10.1037/a0025210

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

Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (Re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 233–240. doi:10.1037/aca0000233

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2014). Reasoned connections: A dual-process perspective on creative thought. *Thinking & Reasoning*, *21*(1), 61-75. doi:10.1080/13546783.2014.895915

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319. doi:10.1037/a0029171

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

Benedek, M., Könen, T., & Neubauer, A. C. (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(3), 273–281. doi:10.1037/a0027059

Briggs, R. O, & Reinig, B. A. (2010) Bounded ideation theory. *Journal of Management Information Systems*, *27*(1), 123-144. doi:10.2753/MIS0742-1222270106

Cheng, L., Hu, W., Jia, X., & Runco, M. A. (2016). The different role of cognitive inhibition in early versus late creative problem finding. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(1), 32–41. doi:10.1037/aca0000036

Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology*, *53*(2), 82–88. doi:10.1037/h0045461

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284–290. doi:10.1037/1040-3590.6.4.284

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millenium. *Gifted Child Quarterly*, *49*(4), 283–356. doi:10.1177/001698620504900402

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, *53*(3), 497–509. doi:10.1037/0022-3514.53.3.497

Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four text-mining methods for measuring elaboration. *The Journal of Creative Behavior*, *55*(2), 517-531. doi:10.1002/jocb.471

Feist, G. J. (1997). Quantity, quality, and depth of research as influences on scientific eminence: Is quantity most important? *Creativity Research Journal*, *10*(4), 325-335. doi:10.1207/s15326934crj1004_4

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*(3), 257-269. doi:10.1080/10400419.2017.1360059

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139. doi:10.1016/j.tsc.2016.12.005

Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 94–112. doi:10.1037/aca0000196

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98*(4), 611–625. doi:10.1111/j.2044-8295.2007.tb00467.x

Goldstein, H. (1991). Nonlinear Multilevel Models, with an Application to Discrete Response Data. *Biometrika*, *78*(1), 45–51. doi:10.1093/biomet/78.1.45.

Gonthier, C., & Roulin, J.-L. (2019). Intra-individual strategy shifts in Raven's matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*. doi:10.1037/xge0000660

Guilford, J.-P. (1950). Creativity. *American Psychologist, 5*, 444-454.

Guilford, J.-P. (1968). *Intelligence, creativity, and their educational implications*. Robert R. Knapp.

Hargreaves, D. J., & Bolton, N. (1972). Selecting creativity tests for use in research. *British Journal of Psychology*, *63*(3), 451–462. https://doi.org/10.1111/j.2044-8295.1972.tb01295.x

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

Heinonen, J., Numminen, J., Hlushchuk, Y., Antell, H., Taatila, V., & Suomala, J. (2016). Default mode and executive networks areas: Association with the serial order in divergent thinking. *PLoS ONE*, *11*(9). doi:10.1371/journal.pone.0162234

Hox, J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ, US: Erlbaum.

Jung, R. E., Wertz, C. J., Meadows, C. A., Ryman, S. G., Vakhtin, A. A., & Flores, R. A. (2015). Quantity yields quality when it comes to creativity: A brain and behavioral test of the equal-odds rule. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.00864

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, *18*, 3-14. doi:10.1207/s15326934crj1801_2

Kim, K. H. (2011). The APA 2009 Division 10 debate: Are the Torrance Tests of Creative Thinking still relevant in the 21st century? *Psychology of Aesthetics, Creativity, and the Arts*, *5*(4), 302-308. doi: 10.1037/a0021917

Kleinkorres, R., Forthmann, B., & Holling, H. (2021). An experimental approach to investigate the involvement of cognitive load in divergent thinking. *Journal of Intelligence*, *9*(1), 3. doi:

Lubart, T., Besançon, M., & Barbot, B. (2011). *Evaluation of Potential for Creativity (EPoC)*. Hogrefe.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46. doi:10.1037/1082-989X.1.1.30

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. doi:10.1037/h0048850

Milgram, R. M., & Rabkin, L. (1980). Developmental test of Mednick's associative hierarchies of original thinking. *Developmental Psychology*, *16*(2), 157–158. doi:10.1037/0012-1649.16.2.157

Mouchiroud, C. and Lubart, T. (2003) Différences intra-individuelles dans le processus de generation d'idées originales chez l'enfant. In Vom Hofe, A., Charvin, H., Bernaud, J. and Guédon, D. (Eds.), *Psychologie différentielle: recherches et réflexions*. Presses Universitaires de Rennes, Rennes, 269–74.

Nijstad, B. A., De Dreu, C. K., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence.

European Review of Social Psychology, *21*(1), 34-77. doi:10.1080/10463281003765323

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts, 8*, 423–432.

Osborn, A. F. (1963). *Applied imagination* (3rd edition). Scribner.

Parnes, S. J. (1961). Effects of extended effort in creative problem solving. *Journal of Educational Psychology*, *52*(3), 117–122. doi:10.1037/h0044650

Phillips, V. K., & Torrance, E. P. (1977). Levels of originality at earlier and later stages of creativity test tasks. *Journal of Creative Behavior*, *11*, 147. doi:10.1002/j.2162-6057.1977.tb00602.x

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Reinig, B. A., Briggs, R. O., & Nunamaker, J. F. Jr. (2007). On the measurement of ideation quality. *Journal of Management Information Systems*, *23*(4), 143-161. doi:10.2753/MIS0742-1222230407

Reinig, B. A., & Briggs, R. O. (2008). On the relationship between idea-quantity and idea-quality during ideation. *Group Decision and Negotiation*, *17*(5), 403-420. doi:10.1007/s10726-008-9105-2

Runco, M. A. (1986). Flexibility and originality in children's divergent thinking. *The Journal of Psychology: Interdisciplinary and Applied*, *120*(4), 345–352. doi:10.1080/00223980.1986.9712632

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment*, *5*(2), 149-156. doi:

Runco, M. A. (1993). Divergent thinking, creativity, and giftedness. *Gifted Child Quarterly*, *37*(1), 16-22.

Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 93-96. doi:10.1037/1931-3896.2.2.93

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. doi:10.1037/0033-2909.86.2.4207

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence*, *40*(4), 343–351. doi:10.1016/j.intell.2012.02.005

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*(2), 79–85. doi:10.1016/j.tsc.2009.06.005

Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2017). Old or new? Evaluating the old/new scoring method for divergent thinking tasks. *The Journal of Creative Behavior*, *51*(3), 216–224. doi:10.1002/jocb.101

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. doi:10.1037/1931-3896.2.2.68

Simonton, D. K. (1977). Creative productivity, age, and stress: A biographical time-series analysis of 10 classical composers. *Journal of Personality and Social Psychology*, *35*(11), 791–804. doi:10.1037/0022-3514.35.11.791

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*(1), 66–89. doi:10.1037/0033-295X.104.1.66

Simpson, R. M. (1922). Creative imagination. *The American Journal of Psychology*, *33*, 234–243. doi:10.2307/1414133

Taylor, C. L., Kaufman, J. C., & Barbot, B. (2021). Measuring creative writing with the storyboard task: the role of effort and story length. *The Journal of Creative Behavior*, *55*(2), 476-488. doi:10.1002/jocb.467

van Rij, J., Wieling, M., Baayen, R., van Rijn, H. (2020). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.4.

Wang, M., Hao, N., Ku, Y., Grabner, R. H., & Fink, A. (2017). Neural correlates of serial order effect in verbal divergent thinking. *Neuropsychologia*, *99*, 92–100. doi:10.1016/j.neuropsychologia.2017.03.001

Ward, W. C. (1969). Rate and uniqueness in children's creative responding. *Child Development*, *40*(3), 869–878. doi:10.2307/1127195

Wood, S. N. (2017) *Generalized Additive Models: An Introduction with R* (2nd edition). Boca Raton, FL, US: Chapman and Hall/CRC.

Zeng, L., Proctor, R. W., & Salvendy, G. (2011). Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity? *Creativity Research Journal*, 23(1), 24–37. doi:10.1080/10400419.2011.545713