How far afield should you go when being creative? Semantic area as a metric of AI's effects on creative ideation

Steven R. Rick[1], **Jennifer L. Heyman**[1], **Pablo Paredes**[2], **Matthew Hong**[2], **Thomas W. Malone**[1]
[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Toyota Research Institute, Los Altos, CA, USA
{srick, heymanj, malone}@mit.edu
{pablo.paredes, matt.hong}@tri.global

Abstract

Generating creative ideas that are both novel and useful is one of the most important goals of any innovation process. But most measures of creativity focus only on the output of a creative process, not on the process itself. Here, we present a new metric for objectively evaluating the creative process itself. The metric, called semantic area, combines semantic embedding models, dimensionality reduction, and 2D area calculations to measure the semantic space covered during a creative task. We then use this method to quantify the relationship between what experimental participants input, see, and select during a creative task and the results of their creative efforts. We find that people who see a more diverse set of ideas (a) select more diverse ideas for their results and (b) have more subjective satisfaction with the results. But, surprisingly, they also select ideas that objective evaluators find less innovative, inspiring, and enjoyable. We close with a discussion of how this method might support interventional approaches that are able to help constructively expand a user's exploration space without distracting them from end goals during a creative process.

Introduction

Creativity as a concept is a complex topic, often referenced when it comes to generating novel ideas or imagining things that do not yet exist. While it is true that creativity is often centered around the idea of originality and novelty when it comes to generating something, it is important to strike a balance between the novelty and the utility of artifacts being created. This is especially critical in spaces where creativity intersects human work. In the work of producing art, creativity tends to intersect with utility by creating surprise or other emotions in the viewer (Tomas 1958) (Tillander 2011) (Filipowicz 2006). In the work of innovation, creativity tends to intersect with utility as it needs to produce something novel that also addresses a real need within an organization or in the marketplace (Amabile 1988) (Yusuf 2009) (Hughes et al. 2018).

It is critical for individuals and organizations to foster and support creativity, but it is often difficult to measure and quantify. For this reason there has been a sizable body of work that has attempted to create tests, scales, and assessment protocols for evaluating creativity (Cropley 2000) (Silvia et al. 2008) (Said-Metwaly, Van den Noortgate, and Kyndt 2017). While these tests are useful for evaluation of the final state of a process, i.e. the final artifact produced, these tests do not give a means for assessing the creative process as a person goes through it. This shortcoming is due in large part to the time consuming nature of evaluation by expert reviewers and the inability for such an approach to scale up very well.

Here we propose a new computational methodology for evaluating the creative process, leveraging the same advances in artificial intelligence and natural language processing that have brought about Large Language Models (LLMs), specifically looking at the embedding models needed to make LLMs work. We situate this work in the domain space of creative product design and show how it not only produces results which link the creative process to the outcome artifact, but also show how it can be used to make the process more transparent, supporting the creation of interventional systems in the future.

Background

Prior work has explored the idea that we might be able to describe creativity through computational means. Much of the earlier work started by trying to classify content as creative or not. Ritchie identified 14 criteria for quantifying novelty (Ritchie 2001) and then later refined them to encompass novelty, quality, and typicality for identifying creative output (Ritchie 2007). Pereira applied Ritchie's criteria in order to quantify novelty of artifacts across evaluation of several systems (Pereira et al. 2005). Peinado et al. focused efforts on evaluating novelty through the lens of reuse (Peinado et al. 2010). Kuznetsova et al. explored this idea further by looking at how language could be modeled as creative or not through the use of lexical composition. Modeling based on divergent thinking techniques, semantic latent space, and affective language allowed them to generate a dataset to classify word pairs through different creative metrics (Kuznetsova, Chen, and Choi 2013). While labeling content as creative or not could be useful, it required reducing ideas to pairs of words. Acar and Runco explored the role of associative distance in a divergent thinking task in order to qualify whether ideas were proximal, i.e. considered remote (distant) vs. close (Acar and Runco 2014). By assessing associative distance against perceptions

of creative attitudes and values, they found remote associations were significantly correlated with creativity, at least in a divergent thinking test. Beketayevab and Runco later explored whether semantic based approaches, in addition to traditional flexibility and originality assessments, could be used to score a divergent thinking test (Beketayev and Runco 2016). They found that semantic flexibility correlated well with traditional flexibility assessments, setting the stage for semantic algorithms to potentially play a role in computational assessment of divergent thinking.

As computational capabilities expanded and became more accessible, prior work began to explore more and more methods for machine learning to play a role in the assessment of creativity. Franceshelli and Musolesi explored the potential for generative deep learning approaches to be used to evaluate facets of creativity such as value, novelty, and surprise (Franceschelli and Musolesi 2022). While this application of deep learning provided an early unsupervised approach (without human involvement), it heavily relied on a training set of data to compare test sets against and tended to restrict wider applicability when working on wide reaching creative tasks. Fan et al. then explored the relationship between creativity in writing tasks and semantic distance (as measured by a trained embedding model using Word2Vec) (Fan et al. 2023). While reliable relationships between semantic distance and creativity were found and the methods allowed for more general applicability that was not overly contrained to a specific training dataset, Word2Vec requires training on a large corpus of content in order to translate textual content into learned word associations (Mikolov et al. 2013). Finally Johnson et al. explored the use of pre-trained embedding models (BERT) to generate semantic distance calculations and evaluate divergent creativity (Johnson et al. 2023). Their use of BERT is the first instance we find of the transformer architecture being used to evaluate creativity, which amplifies the contextual sensitivity of the model due to its self-attention mechanism (Vaswani et al. 2017).

Reflecting upon this prior work, we see two consistent trends: (1) that most methods focus on evaluation of a final artifact, and (2) increasingly large training datasets are necessary to evaluate those artifacts. While these trends are generally fruitful for the field of computational creativity, it is important to find new ways to better support the quantification and qualification of creativity that can consider artifact generation through a sequential process of intermediaries that lead to a final artifact, as well as methods that are not heavily dependent upon large corpora of training data.

Methods

In order to address these gaps identified in prior work, we developed a method for quantifying the 'space' covered during a creative task, a term we name *Semantic Area*. Semantic due to the fact that we leverage embedding models in order to get a contextually aware high-dimensional vector representation of text, and Area as we procedurally reduce that high-dimensional data down to a 2D plane in order to quantify how much space is covered by the set of ideas captured in textual form.

Semantic Area

In order to work with text data, we first cast the corpus of text we want to measure into a high dimensional vector representation. For this we used Google's Universal Sentence Encoder v4 (Cer et al. 2018), but any sufficiently rich embedding model could be used. We selected the Universal Sentence Encoder as it is a pre-trained model, meaning we can use it for inference to convert text into vectors without needing to build an embedding model from our corpus first using something like word2vec. We also chose to use Universal Sentence Encoder as it adjusts the embedding vectors to normalize them based on length of input, allowing for comparison between text strings of different length.

After inferring the high-dimensional vectors of each text string (512-dimensions per text string), we then use Uniform Manifold Approximation and Projection (UMAP) to cluster the data based on cosine similarity and reduce the high dimensional data down to a simpler 2D representation. (McInnes et al. 2018). Specifically, we used the following parameters:

```
UMAP(random_state=12345, n_components=2,
    metric="cosine").fit(vector_list)
```

where `vector_list` contains the array of embedding vectors returned from Universal Sentence Encoder.

UMAP was chosen over other dimensionality reduction methods (like t-SNE) and other clustering methods (like K-means) as UMAP is guaranteed to produce the same dimensionality reduction and clustering, deterministically, if given the same data and parameters. The other methods mentioned have more stochastic behavior leading to random differences between executions, reducing reproducibility. We explicitly set a `random_state` value for this reason, as without that UMAP will employ stochastic optimizations to more quickly cluster and run dimensionality reduction, which we want to avoid for the sake of reproducibility at the cost of a more computationally demanding method.

This dimensionality reduction and clustering allows for the text strings to be represented as points in a 2D plane. While possible to operate in 3D or higher dimensionsion spaces, we chose to work in 2D as this was more intuitive due to the metaphor of 'area' being much simpler than 'volume' or other high dimensional representations. By converting each unique idea or concept in the dataset into a 2D coordinate, we can now begin to analyze semantic area. Using a set of points, for example all the ideas generated by a person or system, we compute a convex hull (the minimal polygon that fits that set of points). We calculate this hull using the quick hull algorithm implementation in SciPy (Barber, Dobkin, and Huhdanpaa 1996) (Virtanen et al. 2020). We then identify the simplicial facets of the convex hull (the points that make up the smallest polygon that contains the set of points), and use the shoelace formula (also known as the Surveyor's Area formula) in order to calculate the area of that polygon (Braden 1986) (Lee and Lim 2017).

The complete Semantic Area calculation pipeline, converting text to embedding vectors, then running dimensionality reduction, and finally calculating a convex hull is rep-

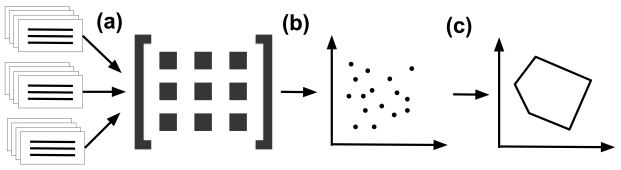


Figure 1: The Semantic Area Pipeline. (a) Text is converted to high-dimensional vectors with an embedding model, (b) the embedding vectors are clustered and dimensionally reduced with UMAP, and (c) a convex hull is calculated to quantify Semantic Area.

resented in Fig 1.

Evaluating Creative Ideation with Semantic Area

In order to use the Semantic Area method to quantify the different effects of ideation interventions on creativity, we obtained a copy of the dataset from the DesignAID study (Cai et al. 2023). We selected this study's data for two core reasons: 1) it was an empirical analysis of people using software to complete a creative task, and 2) the experimental design was a 2x2 factorial where participants were exposed to both a 'direct input mode' and 'new idea mode' during ideation. This meant that the system being used exposed people to both what they directly requested from the system as well as wholly new concepts and ideas based on what they requested from the system. This was of particular interest as we felt the Semantic Area method would be especially useful in capturing the effects of exposure to new ideas on creative task outcomes.

We also received copies of the evaluations that were run, being both subjective and objective evaluations about innovation, fit, inspiration, and enjoyment/liking of the final artifacts produced. Specifically, the participants completed a survey after completing each mood board asking about the extent to which they thought their experiences and final artifacts were innovative (1-5, $M = 3.40, SD = 1.05$), fitting of the prompt of "a chair for children" (1-5, $M = 3.90, SD = 1.11$), inspirational (1-7, $M = 5.01, SD = 1.59$), and enjoyable (1-7, $M = 5.27, SD = 1.75$). Similarly, after all mood boards were created, they were presented to a separate group of evaluators who rated those artifacts based on their level of innovation (1-5, $M = 4.19, SD = 0.88$), fit (1-5, $M = 4.34, SD = 1.32$), inspiration (1-7, $M = 3.65, SD = 0.77$), and liking (1-7, $M = 4.00, SD = 0.77$).

Data preparation

As the Semantic Area method is centered around two-dimensional area calculations, we first subset the DesignAID dataset for users who provided at least 3 unique inputs in order to have a minimum of 3 points available for running area calculations (N = 111, from the original 115). From these 111 users, 9106 text strings were collected across the input text, seen text, and selected text. 5058 of those were unique, with 1093 coming from users in the Direct Input mode and 3965 coming from users in the Diverse (New Idea) mode, with 80 overlapping (found in both sets). For simplicity we will refer to direct input mode as Direct and new idea mode as Diverse. After the data was clustered and dimensionally reduced by UMAP, we had 2D coordinates available for each unique string in the dataset. A 2D representation of the data is shown in Fig 2.

Calculating and Using Semantic Area

After reducing the textual data to a 2D coordinate space, we explore the power of Semantic Area to describe a person's creative journey. By selecting the set of texts that a user input to the system, texts that a user saw from the system, and texts that a user selected from the system, we can visualize the area covered by these three categories of data. An example of this is shown in Fig 3. When a user was in Direct input mode they would have been restricted to only see and select from the text that they input, as no new ideas would have been created by the system. When in New Idea mode their input would have directed the system to generate new ideas with similar context to their input, meaning they would see both their own inputs and the new ideas generated by the system. It is important to note that Seen always encapsulates both Input and Selected as the superset of all texts.

While running the Semantic Area calculations across all data for a user can show the total Semantic Area, we can also measure and visualize how that area changes over time. We do this by procedurally adding one new point at a time based on the temporal sequence in which a user encountered the idea. We recompute the convex hull at each time step to determine how the semantic area changes with each new idea. Fig 4 shows this with the same example data that was used in Fig 3, showing how the *Seen* area changes with each new text.

Results

Having now introduced the Semantic Area methodology and demonstrated examples of what the method can enable us to see, we scale up our analysis to report on the relationships found in the DesignAID dataset. As a reminder, the DesignAID study instructed individuals to create two moodboards depicting a "chair for children" using text input to generate images. All participants were exposed to two treatments: the Diverse (New Idea) mode where the input they provided was semantically expanded by a Large Language Model, and the Direct input mode where the input they provided remained unchanged. Those texts were then used to create images and participants designed their final moodboards by selecting from the images they produced during the task.

Semantic Area and Mode of Interaction

Before conducting any analysis, we pre-registered a hypothesis on OSF[1]. In order to understand the relationship between the texts a user input, saw, and selected from across the experimental treatments, we plot each user's data, showing how the semantic area of input leads to the semantic area

[1]https://doi.org/10.17605/OSF.IO/KM2ZQ

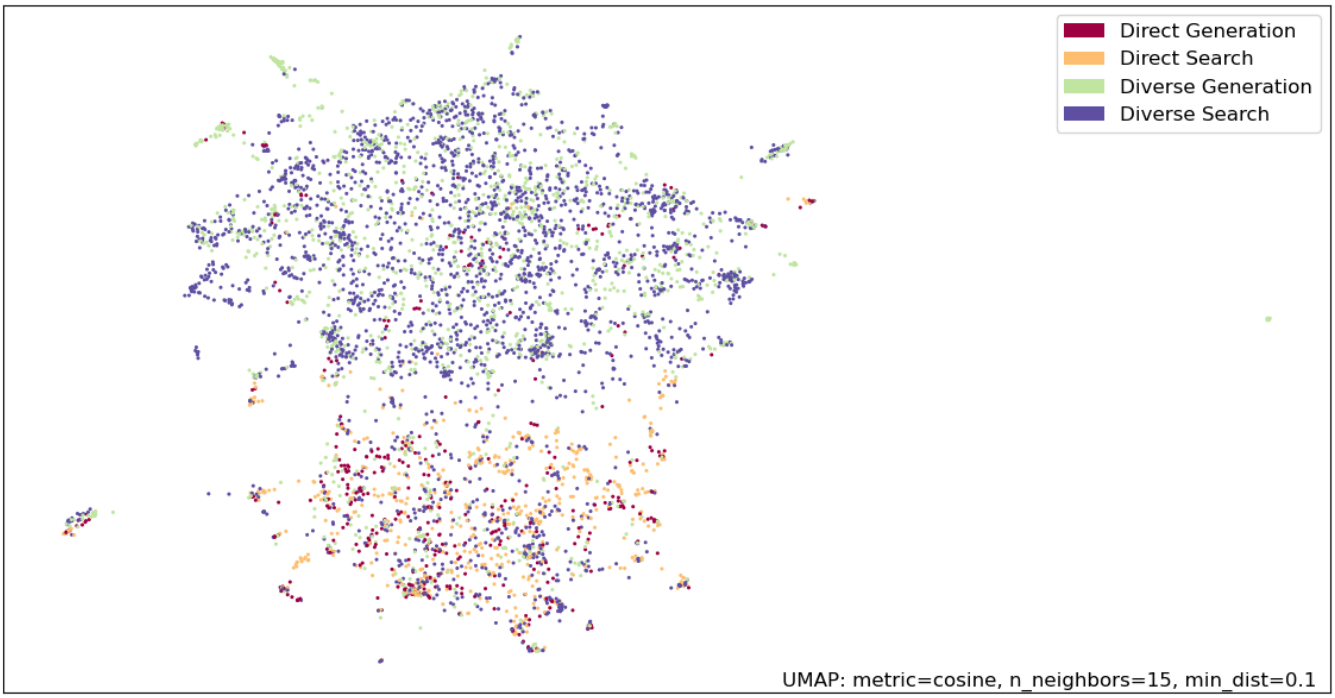


Figure 2: The UMAP dimensionality reduction and 2D clustering of the text strings from the DesignAID dataset

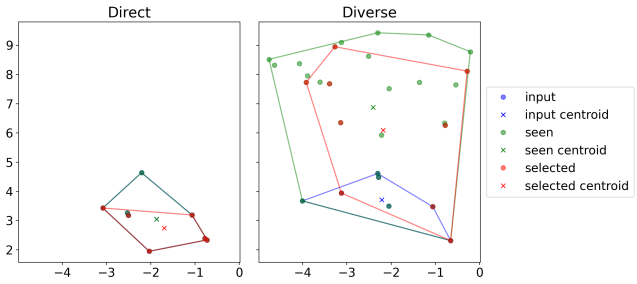


Figure 3: An example of a user's Semantic Area across what they input, what they saw, and what they selected. The Direct input mode is on the left, and Diverse (New Idea) mode is on the right. All coordinates of all ideas are plotted as points, the convex hull is computed and drawn with lines, and the centroid is calculated and shown with an X.

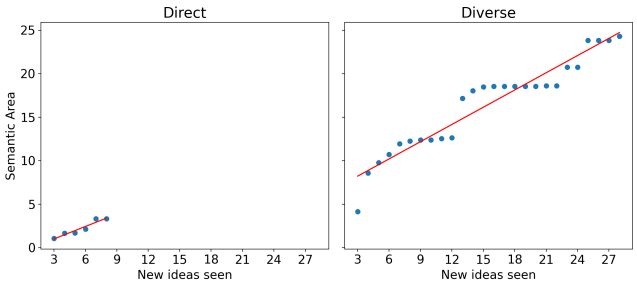


Figure 4: An example of a user's Semantic Area changing over time. Direct input mode is on the left, and Diverse (New Idea) mode is on the right. Each point is the area at that step in the creative process (adding new 'Seen' ideas). The red line is the least squares regression of the data (line of best fit).

of seen ideas and finally to the semantic area of selected ideas. This is shown in aggregate in Fig 5. It is worth noting that while users all started from different input areas, those in the Diverse condition (where they were exposed to New Ideas) saw area values increase when transitioning from input to seen ($M_{Diff} = 21.16$), whereas those in the Direct input conditions did not see a significant change between input and seen areas ($M_{Diff} = .02$). This makes intuitive sense given that the Diverse intervention was explicitly trying to expose the end user to a wider range of ideas. A first finding surfaced that, on average, users in the Diverse states selected items that produced a larger semantic area than the set of inputs they provided ($M_{Diff} = 5.63$). Users in the Direct input conditions always selected ideas that produced a smaller semantic area than the area of the inputs they had provided ($M_{Diff} = -1.72$). This is shown by the average value of each condition (shown as a red line on the charts).

We found that there were significant differences between the Diverse input conditions and the Direct input conditions when examining input area ($t = -2.04$, $p = .04$, $M_{Diff} = -1.32$), seen area ($t = -19.44$, $p < .0001$, $M_{Diff} = -22.63$) and selected area ($t = -8.54$, $p < .0001$, $M_{Diff} = -8.54$). Specifically, the semantic areas for the Diverse conditions were significantly greater than the semantic areas for the Direct conditions (Fig 6).

This directly confirms the hypothesis we originally regis-

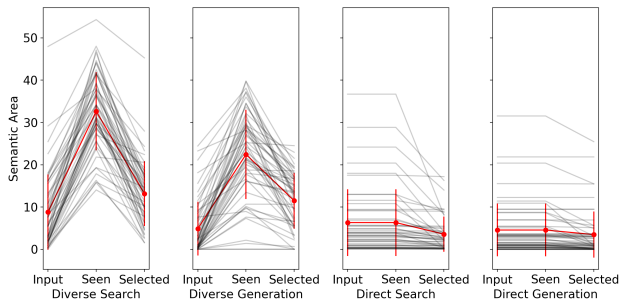


Figure 5: Aggregate chart plotting the input area, seen area, and selected area for each participant grouped by treatment condition. Those in the Diverse conditions always increased from input to seen. They also, on average, increased from input to selected. Those in the Direct conditions always decreased area from input to selected. Average values are shown with red lines.

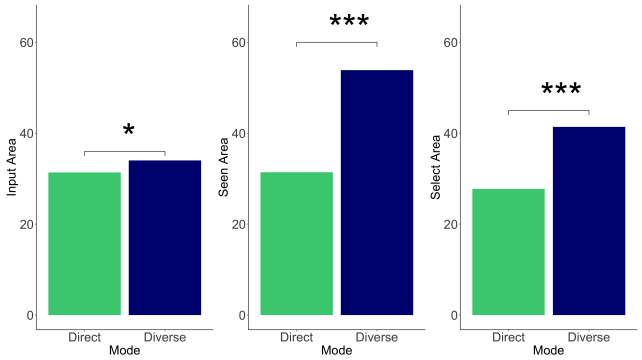


Figure 6: Within-Subjects T-Test Examining the Differences Between Input (left), Seen (middle), and Select (right) Semantic Areas Between Low (green) and High (blue) Diversity conditions

tered and generates new questions about whether there was any relationship between Semantic Area (input, seen, and/or selected) and the subjective and objective assessments run by Cai, et al. We expand our analysis to look at the interplay between these data.

Modeling the Creative Process Over Time With Semantic Area

We begin to get a glimpse into the creative process by looking at how the semantic area changes over time as users explored an idea space and/or are given exposure to a greater number of new ideas. An example of this is shown in Fig 4.

We found that as a person explores more ideas, the semantic area increases ($b = 0.34$, $z = 57.95$, $p < .0001$). Importantly, this effect is stronger in the Diverse New Idea mode ($b = 0.29$, $z = 53.25$, $p < .0001$), although it is still significantly positive in the Direct Input mode ($b = 0.23$, $z = 12.08$, $p < .0001$).

	b	se	p
Input Area			
Subjective Measures			
Innovation	-0.02	.01	.16
Fit	0.02	.01	.17
Inspiration	-0.01	.02	.54
Enjoyment	-0.05	.02	.01*
Objective Measures			
Innovation	0.01	.01	.28
Fit	0.00	.01	.79
Inspiration	0.01	.01	.12
Liking	0.01	.01	.29
Seen Area			
Subjective Measures			
Innovation	-0.02	.01	.003**
Fit	-0.01	.01	.05*
Inspiration	-0.02	.01	.005**
Enjoyment	-0.05	.01	<.0001***
Objective Measures			
Innovation	-0.00	.00	.35
Fit	-0.01	.01	.10
Inspiration	-0.00	.00	.51
Liking	-0.00	.00	.28
Select Area			
Subjective Measures			
Innovation	-0.01	.01	.62
Fit	0.01	.01	.49
Inspiration	-0.01	.02	.64
Enjoyment	-0.04	.02	.02*
Objective Measures			
Innovation	-0.01	.01	.13
Fit	-0.02	.01	.07
Inspiration	-0.01	.01	.21
Liking	-0.01	.01	.10

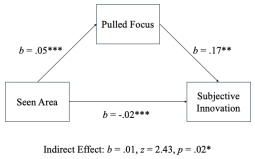
Table 1: Semantic Area Predicting Subjective and Objective Measures. Seen area had significant effect on all subjective measures, while input area and select area had significant effects on subjective enjoyment. No significance was noted between areas and objective measures. *$p < .05$, **$p < .01$, ***$p < .001$

Semantic Area and Evaluated Outcomes

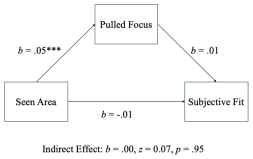
We conducted a series of linear regressions to examine whether the semantic area of texts that participants input, saw, or selected predicted subjective and objective measures of innovation, fit to the originally prompt, inspiration, and enjoyment/liking. Table 1 showcases the results of our analysis.

Generally, neither input area nor selected area were significantly associated with objective or subjective measures (all $ps > .07$). The exception to this was that both the input ($b = -.05$, $z = -2.50$, $p = .01$) and selected ($b = -.04$, $z = -2.37$, $p = .02$) areas were significantly negatively associated with subjective ratings of enjoyment. In other words, those who input a wider range of text prompts tended to rate their experience creating the mood boards as being less enjoyable.

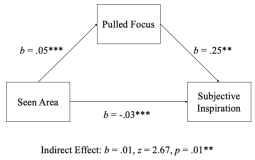
When examining the association between the Seen area and the objective and subjective measures, we found that the Seen area was not significantly associated with any of the objective measures (all $ps > .1$). However, the Seen Area was significantly negatively associated with all of the subjective measures (all $ps < .05$). That is, those who saw a wider range of outputs from the system tended to rate their experiences as being less innovative, less fitting to the prompt, less inspiring, and less enjoyable.



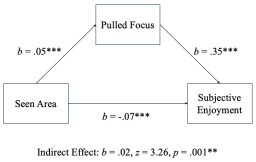
(a) Mediation Between Seen Area, Pulled Focus, and Subjective Innovation



(b) Mediation Between Seen Area, Pulled Focus, and Subjective Fit

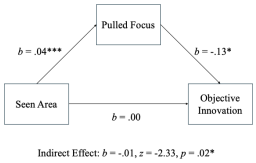


(c) Mediation Between Seen Area, Pulled Focus, and Subjective Inspiration

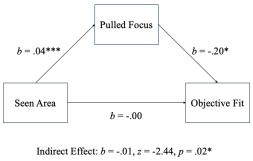


(d) Mediation Between Seen Area, Pulled Focus, and Subjective Enjoyment

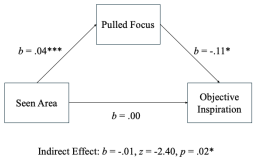
Figure 7: Mediation Between Seen Area, Pulled Focus, and Subjective Measures



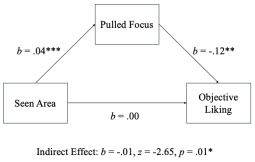
(a) Mediation Between Seen Area, Pulled Focus, and Objective Innovation



(b) Mediation Between Seen Area, Pulled Focus, and Objective Fit



(c) Mediation Between Seen Area, Pulled Focus, and Objective Inspiration



(d) Mediation Between Seen Area, Pulled Focus, and Objective Like

Figure 8: Mediation Between Seen Area, Pulled Focus, and Objective Measures

Seeing New Ideas Pulls Focus and Influences What You Select

To examine whether the ideas that one sees influences the final outcome more than the idea that one originally inputs, we calculated the distance between the centroids of the input, seen, and selected ideas. By centroid of the ideas, we mean the average value of all coordinates per category, not just the coordinates that make up the convex hull. We then conducted t-tests to examine whether the distance between the Input and Seen centroids was significantly different from the distance between the Seen and Selected centroids. If the original ideas that one inputs are more strongly associated with the final outcome, we reason that we would see a smaller distance between the Input centroid and Selected centroid as compared to the distance between the Seen centroid and Selected centroid. In contrast, if the ideas that one sees has a stronger influence on the final selected ideas, we would see the distance between the Seen and Selected centroids as being smaller than the distance between the Input and Selected centroids.

Indeed, we found that the average distance between the Seen and Selected centroids ($M = 0.82$) was significantly smaller than the distance between the Input and Seen centroids ($M = 1.53$, $p < .0001$), indicating that the ideas that one sees tend to have a stronger influence on the final output than the ideas that one originally inputs.

This implies that the ideas that one sees may tend to *pull focus* away from the ideas that one originally inputs, thus encouraging a larger semantic area of exploration to consider when selecting content for the final output. This may also have other effects upon the creative process, possibly enhancing the innovativeness of a final artifact by encouraging a wider set of concepts to be considered or, inversely, distracting the focus of a person and bringing them further away from their initial target goal.

Effects Between What People Do, What People See, and What People Select

To explore the effects of this *pulled focus* on creative process outcomes, we conducted several mediation analyses to examine the causal effect between Seen area, pulled focus, and both subjective and objective measures. That is, does greater Seen area predict greater distance between the Input and Selected centroids which, in turn, predicts subjective and objective measures related to final artifact outcomes?

When examining the subjective measures, we find a significant indirect effect between Seen area, pulled focus, and innovation ($b = .01$, $z = 2.43$, $p = .02$, Fig 7a), inspiration ($b = .01$, $z = 2.67$, $p = .01$, Fig 7c), and enjoyment ($b = .02$, $z = 3.36$, $p = .001$, Fig 7d). In other words, those who saw a wider breadth of ideas tended to select ideas that were semantically further from their original inputs which, in turn, predicted higher levels of subjective innovation, inspiration, and enjoyment. The indirect effect for subjective ratings of fit was not significant ($b = .01$, $z = 0.07$, $p = .95$, Fig 7b).

Interestingly, when we examine the objective measures, we find the opposite effect. Specifically, those who saw a wider breadth of ideas tended to select ideas that were semantically further from their original inputs, which then predicted significantly lower levels of objective innovation ($b = -.01$, $z = -2.33$, $p = .02$, Fig 8a), fit ($b = -.01$, $z = -2.44$, $p = .02$, Fig 8b), inspiration ($b = -.01$, $z = -2.40$, $p = .02$, Fig 8c), and liking ($b = -.01$, $z = -2.65$, $p = .01$, Fig 8d).

In sum, seeing a wider breadth of ideas led to selecting outputs that were significantly different from the originally inputs. This, in turn, led to a more positive subjective experience for those creating the mood boards, but a more negative objective outcome when evaluating those mood boards.

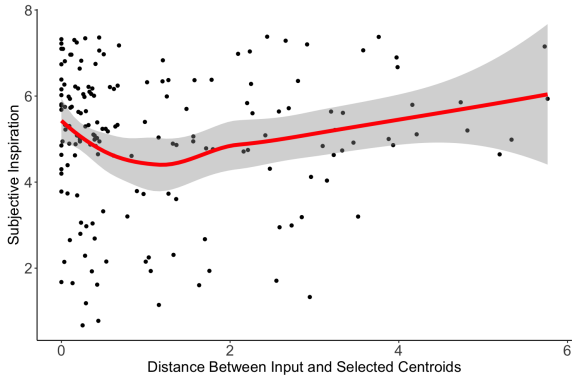


Figure 9: Non-Linear Association Between Input and Selected Centroid Distance and Subjective Inspiration

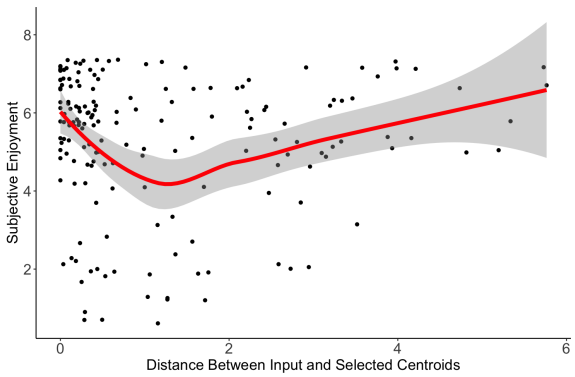


Figure 10: Non-Linear Association Between Input and Selected Centroid Distance and Subjective Enjoyment

Is There An Optimal Level of New Idea Exposure?

Given these rather surprising findings (that exposure to more new ideas produced subjective improvement and objective detriment), we asked whether there is an optimal level of new ideas to be exposed to in order to provide the most positive subjective experiences and objective outcomes. To do this, we conducted a series of quadratic regressions to examine if there is a non-linear relationship between both Seen area and Centroid Distance and our subjective and objective measures.

We found a significant quadratic association between the distance between the Input and Selected centroids when predicting both subjective inspiration ($b = .12$, $z = 2.08$, $p = .04$, Fig 9) and enjoyment ($b = .21$, $z = 3.50$, $p = .001$, Fig 10). That is, when the distance between the Input and Selected centroids is low, there is a marginal negative association with subjective inspiration ($b = -0.48$, $z = -1.72$, $p = .09$). In contrast, when the distance between the centroids is high, there is a marginal positive association with inspiration ($b = 0.17$, $z = 1.71$, $p = .09$). Similarly, there is a significant negative association between the centroid distance and subjective enjoyment when the distance is low ($b = -0.97$, $z = -3.27$, $p = .001$), although this association becomes marginally positive when the distance between the centroids is greater ($b = .19$, $z = 1.77$, $p = .09$). No other quadratic associations were significant, either when examining subjective measures ($p > .1$) or objective measures ($p > .1$).

This indicates that the benefits of diverse new ideas are highest when the selected ideas are particularly different from the originally input ideas. However, the benefits may not be as robust if the selected ideas are only moderately different from the originally input ideas. In other words, in order to reap the benefits of semantic area exploration subjectively, one must step far outside of their comfort zone and explore ideas far from where they originally started. Given our analysis, further work is required to really understand the effect of new idea exposure on objective outcomes.

Discussion

Reflecting upon the semantic area methods we developed and results of our analysis using it to quantify a creative ideation task, we find a number of surprising results. We noted that people who put more diverse texts into the system had a worse subjective experience, likely due to the higher effort required to produce new ideas from scratch. We also noted that exposure to a more diverse set of ideas can enhance subjective perception but reduces objective performance, likely due to new ideas pulling focus away from the starting location.

What You See Is What You Think

We found that exposure to a wider range of ideas (that are semantically further from one another) has mixed effects on outcomes drives forward a consideration that what you *see* influences what you *think*. In this manner, exposure to new perspectives, concepts, and ideas can broaden the creative search space. This in turn can drive up a person's feelings of innovation. At the same time, this can *pull focus* and shift what a person is attending to, leading to not staying on target within a task and producing objectively worse outcomes.

Considering the growing presence of generative AI technologies and the ability of the technology to generate seemingly endless new content, it is important we consider how these new computational abilities are designed and used to empower people. There is potential to augment with inspiration from new ideas, but also risk of distraction or overloading with too many new perspectives.

Semantic Area and Interventional Processes

Keeping the study we got our data from in mind, users were left in a perpetual state of always getting new ideas or never getting anything novel. While informative for the sake of empirical evaluation, our methods highlight the tricky balance to be struck between enough new ideas and too many. Future work with generative AI creating new ideas should consider more human autonomy in the process. By enabling users to decide when they do or do not want new ideas, they can take control of this process and steer the technology to their benefit.

However, even this requires people to remain aware of their state in the creative process. We believe the Semantic Area method for modeling exposure to ideas over time might present a powerful new way to build dynamic creative interventions. By detecting when people are not exploring and encouraging it, or noticing when people are losing focus and pulling them back to a focal point, future systems might be able to drive people toward optimal outcomes without them needing to stop and determine the next best step. Semantic Area seems especially useful for ideation process recommendation, presenting new ways to quantify and qualify activities as divergent or convergent based on what someone has done over time, and then recommending potential next steps.

Limitations and Future Work

We acknowledge there are some limitations of the methodology and analysis. First, Semantic Area is a relative calculation, meaning the area comparisons need to be from data that has had their embedding vectors come from the same embedding model, and their clustering and dimensionality reduction need to use the same corpus of data. Because it is a relative evaluation, it is important to not consider these Semantic Area values as global measures of creativity, but rather localized measures best suited to comparison of data collected from common tasks on common themes, or for the evaluation of an individual over time.

Since the Semantic Area method uses clustering and dimensionality reduction this also means that without a starting corpora of text data, the initial clustering and reduction from embedding vectors will move around dramatically as data is collected until a somewhat steady state is reached with a larger corpus. As such, the method currently proves to be most useful in post-hoc analysis. Future work could adapt the approach to do away with dimensionality reduction as methods for higher dimensional convex hulls and volumetric proxies do exist, but at the cost of a loss of intuitive understanding of the computed value or straightforward visualization of the data.

Another important limitation of the analyses is that we did not get a comprehensive understanding of the creative process as a whole from the dataset we had access to. That is, the data we analyzed focused on objectively assessing the final artifacts produced and the overall experience reported by subjective and objective measures of innovation, fit, inspiration, and enjoyment/liking. All of this was collected after the creative process had concluded. While we can confidently report on end outcomes, it is less clear how our results fit a proper characterization of the process over time. Future work should explore more process oriented evaluations (with something like frequent checkpoint surveys) in order to connect Semantic Area more tightly with the creative journey rather than only the final artifact produced.

Conclusion

In conclusion we present Semantic Area, a new method for quantifying creativity from text embeddings, clustering, and convex hull calculations. We showcase the utility for this method to describe not just the final state a user arrives at through a creative task, but also how it might be used to characterize the process over time during a creative journey. Using data collected from a study where users were exposed to new ideas to increase inspiration and innovation, we highlight the mixed subjective and objective effects of new ideas generated by AI on creative outcomes. Specifically we show that new idea exposure can be beneficial to subjective inspiration and detrimental to objective fit on final outcomes, likely due to pulled focus. We also show that new idea exposure can be detrimental to subjective enjoyment, likely due to increased cognitive effort evaluating new concepts. New ideas can be helpful, but too many can get in the way.

We see this work as surfacing new considerations around the effects of AI on the creative process. We highlight opportunities for computational creativity at large to better design systems that support people doing creative work with AI. Augmenting how people and computers do creative work together requires us to build systems that can make sense of the creative journey and intervene in human creative work to promote innovation and inspiration without overwhelming people or causing them to lose focus.

Author Contributions

Author 1 led the development of the methodology and analysis. Author 2 helped with statistical analysis. Authors 3, 4, and 5 all helped with interpretation of results as well as research plan development. All authors wrote and reviewed the manuscript.

Acknowledgments

This research was supported in part by the Toyota Research Institute and also as part of the Collective Intelligence Mission of the MIT Quest for Intelligence.

References

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

Amabile, T. M. 1988. A model of creativity and innovation in organizations. *Research in organizational behavior* 10.

Barber, C. B.; Dobkin, D. P.; and Huhdanpaa, H. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* 22(4):469–483.

Beketayev, K., and Runco, M. A. 2016. Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's journal of psychology* 12(2):210.

Braden, B. 1986. The surveyor's area formula. *The College Mathematics Journal* 17(4):326–337.

Cai, A.; Rick, S. R.; Heyman, J. L.; Zhang, Y.; Filipowicz, A.; Hong, M.; Klenk, M.; and Malone, T. 2023. Designaid: Using generative ai and semantic diversity for design inspiration. In *Proceedings of The ACM Collective Intelligence Conference*, 1–11.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar,

C.; et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 169–174.

Cropley, A. J. 2000. Defining and measuring creativity: Are creativity tests worth using? *Roeper review* 23(2):72–79.

Fan, L.; Zhuang, K.; Wang, X.; Zhang, J.; Liu, C.; Gu, J.; and Qiu, J. 2023. Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology* 60(5):e14239.

Filipowicz, A. 2006. From positive affect to creativity: The surprising role of surprise. *Creativity Research Journal* 18(2):141–152.

Franceschelli, G., and Musolesi, M. 2022. Deepcreativity: measuring creativity with deep learning techniques. *Intelligenza Artificiale* 16(2):151–163.

Hughes, D. J.; Lee, A.; Tian, A. W.; Newman, A.; and Legood, A. 2018. Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly* 29(5):549–569.

Johnson, D. R.; Kaufman, J. C.; Baker, B. S.; Patterson, J. D.; Barbot, B.; Green, A. E.; van Hell, J.; Kennedy, E.; Sullivan, G. F.; Taylor, C. L.; et al. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods* 55(7):3726–3759.

Kuznetsova, P.; Chen, J.; and Choi, Y. 2013. Understanding and quantifying creativity in lexical composition. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1246–1258.

Lee, Y., and Lim, W. 2017. Shoelace formula: Connecting the area of a polygon and the vector cross product. *The Mathematics Teacher* 110(8):631–636.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* 3(29):861.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26.

Peinado, F.; Francisco, V.; Hervás, R.; and Gervás, P. 2010. Assessing the novelty of computer-generated narratives using empirical metrics. *Minds and Machines* 20:565–588.

Pereira, F. C.; Mendes, M.; Gervás, P.; and Cardoso, A. 2005. Experiments with assessment of creative systems: an application of ritchie's criteria. In *Proceedings of the workshop on computational creativity, 19th international joint conference on artificial intelligence*, volume 5, 05.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB'01 Symposium*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Said-Metwaly, S.; Van den Noortgate, W.; and Kyndt, E. 2017. Methodological issues in measuring creativity: A systematic literature review. *Creativity. Theories–Research-Applications* 4(2):276–301.

Silvia, P. J.; Winterstein, B. P.; Willse, J. T.; Barona, C. M.; Cram, J. T.; Hess, K. I.; Martinez, J. L.; and 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.

Tillander, M. 2011. Creativity, technology, art, and pedagogical practices. *Art Education* 64(1):40–46.

Tomas, V. 1958. Creativity in art. *The Philosophical Review* 67(1):1–15.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods* 17(3):261–272.

Yusuf, S. 2009. From creativity to innovation. *Technology in society* 31(1):1–8.