

**Back to the Basics: Correspondences Between Creativity Measures and Creative Output  
(Painting)**

Anna P. Smith<sup>a</sup>, Nathaniel Barr<sup>b</sup>, Alexander Christensen<sup>c</sup>, Chloe Williams<sup>a</sup>, Jonathan Schooler<sup>d</sup>,  
Anjan Chatterjee<sup>c</sup>, Paul Seli<sup>a</sup>

<sup>a</sup>Department of Psychology and Neuroscience, Duke University, Durham NC, USA

<sup>b</sup>Faculty of Humanities and Social Sciences, Sheridan College, Oakville, ON, Canada

<sup>c</sup>University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

<sup>d</sup>University of California, Santa Barbara, Santa Barbara, CA, USA

**Author Note**

Correspondence should be addressed to Anna P. Smith, Department of Psychology & Neuroscience, 417 Chapel Drive, Duke University, Durham, NC 27708, USA. E-mail: [anna.p.smith@duke.edu](mailto:anna.p.smith@duke.edu).

Word Count: 4986

**Abstract**

Researchers since the 1950s have invested a great deal in creating “gold standard” creativity assessments that can be administered in a controlled laboratory setting. Despite the successes in developing reliable and widely used instruments, these efforts have come at the cost of not using ecologically- and face-valid tasks. In this paper, we describe a novel protocol to fill this critical gap by bringing participants into the laboratory to paint. We also gave them common laboratory-based creativity tasks and questionnaires. First, we were interested in the feasibility of this method: could participants follow instructions to produce paintings of variable quality, as well as whether independent raters would agree on judgments of painting quality. Second, we were interested in the domain-specificity of abstract painting, and whether performance on either or both the Alternate Uses Task and the Test of Creative Imagery Abilities correlated with painting quality. Finally, we examined whether painting quality was associated with particular personality traits, mindsets towards creative activities, fluid intelligence scores, and existing creative hobbies and achievements. Our findings suggest that an in-lab painting task is both feasible and informative, and may help separate creativity in the visual arts domain from verbal creativity and general intelligence.

*Keywords:* creativity; divergent thinking; fluid intelligence; painting

**Back to the Basics: Correspondences Between Creativity Measures and Creative Output  
(Painting)**

Creativity has stood as a central element of what it is to be human for millennia. From cave art in Sulawesi that was painted over 40,000 years ago (Aubert, et al., 2019) to the recent development of a scalable pipeline for developing completely biological machines (Kriegman, Blackiston, Levin & Bongard, 2020), and innumerable artistic, technological, scientific, and social innovations in between, the capacity for creativity has shaped our experiences and afforded us new possibilities.

Despite this centrality throughout the ages, the formal study and measurement of creativity is a recent development. Tracing the origins of modern empirical approaches to measuring creativity takes us back less than a century. Many psychologists returned from their military posts after World War II with a desire to rectify the shortcomings of existing approaches to understanding human abilities and to meet the practical need to select individuals for positions requiring extraordinary levels of ingenuity and problem-solving abilities. While intelligence was clearly an important metric of human ability, capacity for creativity was not targeted. J.P. Guilford's (1950) APA address, which challenged the long-held assumption that creativity and intelligence were largely synonymous, identified a glaring dearth of research exploring the topic, and emphasized the need for the field to take seriously the question of testing creative ability through valid and reliable instruments (see Ausubel, 1964).

This challenge yielded many successful tools, including the Torrance Tests of Creative Thinking (Torrance, 1966), which includes tests of verbal and figural creativity, and the Alternate Uses Task (Guilford, 1967), to measure semantically-divergent thinking. While these tests are often considered "gold standards" in creativity assessment, vestiges of the historical

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

problem that inspired their development remain. Frank Barron (1965) described the problems of aptitude test construction, citing the practical demand for assessments which could be administered to groups rather individuals and could be scored without a rater, and which have simple scoring metrics amenable to plotting frequency distributions that could be compared. He went on to note that this “set of requirements, however, immediately bumps head-on to the nature of the creative act, which commonly is quite complex and to be recognized must have an observer capable of embracing its complexity” (Barron, 1965, p. 13).

Such observations cut to the core of a major concern of psychometric assessments in measuring creativity. The AUT, positions itself as a test of divergent thinking ability s. Performance on the AUT is determined by participants’ ability to produce “novel and useful” uses for everyday objects, such as “brick” or “marble.” The scoring criteria are credited to James Guilford in 1950, who defined creativity as “uncommon, yet acceptable responses to items,” and later more clearly by Morris Stein in 1953 as “a novel work that is accepted as tenable and useful” (Runco & Jaeger, 2012). Although divergent thinking is not strictly synonymous with creativity, it is not unusual for researchers to characterize it as an acceptable indicator of creative potential, based on its correspondence to real-world creative activities and achievements (Beaty et al., 2018; Jauk, Benedek, & Neubauer, 2014; Runco & Acar, 2012).

For all of the predictive power of AUT, one elephant remains in the room: the test does not resemble everyday instances of creativity, such as culinary experimentation, landscaping, web design, or songwriting, to use examples from the Creative Achievement Questionnaire (Carson, Peterson, & Higgins, 2005). One possibility is that the AUT, and other laboratory measures of divergent thinking, measure the same, or a closely overlapping, construct as do intelligence test. Studies report that scores on general and fluid intelligence tests account for

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

between 24 and 63% of variance on creativity, depending on how they are measured (Frith et al., 2021; Silvia & Beaty, 2012). Such results suggest that perhaps we have yet to achieve Guilford's original call to develop measures of that disentangle creativity from intelligence. Despite decades of research, little is known of how similarly individuals perform on standard creativity tasks in relation to more hands-on self-evident creativity performances.

Why the field has focused on these tasks despite their shortcomings is a matter of speculation. Perhaps the same task constraints that facilitated military personnel selection remain attractive to those seeking easy-to-administer assessments within laboratory environments (see McNemar, 1946; Rubenstein, 1982; Kimmel, 1996). Perhaps, as Hintzman (2011) argued was the case in the field of learning and memory, researchers have become overly task-focused and would benefit from taking a broader view. Maybe the field became drawn into dispelling myths that creativity resides solely within the arts or is too ephemeral to measure (e.g., Benedek et al., 2021); that we have clung to tasks divorced from lay conceptions of creativity. Whatever the case, while value remains in tasks such as the AUT, the constraints on our measurement tools have moved from practical to self-imposed. The time is now to experiment with more naturalistic and face-valid measures of creativity.

Some examples of more ecologically-valid creativity tasks have been examined in the literature, though they not widely administered. Humor production—the ability to be funny on the spot—has been studied using cartoon captions (e.g., New Yorker cartoons), odd noun-noun combinations (e.g., yoga bank), and joke stems completion tasks, such as completing a joke about a friend who sings horribly (Christensen et al., 2018; Nusbaum et al., 2017). Metaphor production has featured in similar tasks, such as simile and figurative statement-completion (Beaty & Silvia, 2013; Christensen & Guilford, 1963; Silvia & Beaty, 2012). Recent

breakthroughs in machine learning have also enabled researchers to automatically score creative writing (Johnson et al., 2022). A limitation of these tasks is that they all heavily rely on verbal and written ability and therefore only capture a specific application of creativity.

Outside of the verbal domain, some researchers have examined musical improvisation—from jazz pianists to classical musicians and freestyle rappers—to assess domains of non-verbal creativity (Beaty, 2015). More commonly, researchers have used drawing tasks such as the Test for Creative Thinking-Drawing Production (TCT-DP; Jellen & Urban, 1989) and Test of Creative Imagery Abilities (TCIA; Jankowska & Karwowski, 2016) to assess non-verbal creativity. The Torrance Tests of Creative Thinking (Torrance, 1966), from which the AUT originated, also used drawings as a part of its battery. These drawings tasks are usually pen-and-paper and provide incomplete figures or lines that the participants expand on. Although a step forward in the measurement validity of creative thinking, these tasks constrain participants to two hues (grey of the pencil and white of the page) and do not allow freedom of expression, as the incomplete figures or lines must be used and are often incorporated into the scoring of the task. Taken together, these more ecologically valid creativity tasks are limited to the verbal domain or in the scope of the participant’s creative expression.

Thus, here, we report on a novel protocol that fills that critical gap: to examine the relationship between widely-used laboratory creativity measures and a natural expression of creativity. To do so, we brought participants into the laboratory to paint freely. We investigated the relationships between the quality of freeform abstract paintings and commonly collected laboratory-based creativity tasks and questionnaires. First, we wished to determine the feasibility of this method: whether participants could follow instructions to produce paintings of systematically variable quality, as well as whether independent raters could agree on a judgment

of painting quality with acceptable reliability. Second, we were interested in the domain-specificity of abstract painting, and whether either or both the AUT (Guilford, 1967) and the TCIA (Jankowska & Karwowski, 2016) correlated with painting quality. Finally, we examined whether painting quality was associated with various common inclusions in creativity research, including personality and mindset measures.

## **Method**

### **Participants**

This study was approved by the Institutional Review Board of Duke University. 100 participants were recruited from an undergraduate online subject pool and were offered course credit in return for one-hour participation in this study. Individuals were eligible if they were over the age of 18 and spoke English fluently. The data from one participant was excluded from analysis because they did not follow directions, bringing the total sample size to 99 (mean age = 19.13, SD = 1.08, 31% male).

Participants were told at sign-up that they would be given a number of questionnaires that would assess their personality, fluid intelligence, creative mindset, and creative history. They would also complete three creativity tasks where they would be asked to perform activities such as completing a picture with the shape given or devising novel uses for everyday objects. Finally, they would be asked to complete an abstract painting using the materials provided. They were told that the study would take approximately 1 hour, and that they would receive one credit hour at completion.

### **Materials**

Instrument text is included in Appendix B.

*Primary Creativity Measures of Interest*

We administered three tasks to measure in-lab creative performance: a classic laboratory-based task in the verbal domain (AUT), a laboratory task in the visuospatial domain (TCIA), and an ecologically-valid opportunity to showcase visual artistic ability (painting). To estimate the quality of responses on these measures, we used many-facet Rasch models (MFRM; Linacre, 1994) on the ratings of three trained experimenters.

Recent work using creativity tasks has demonstrated that MFRM are effective for correcting rater biases (e.g., being more severe or lenient than other raters; Primi et al., 2019). Like single-facet Rasch models for ability tests, many-facet Rasch models adjust ability estimates for the lenience (or severity) of the raters but also the difficulty of the prompt (e.g., “pen” may be more difficult to come up with alternatives uses for than a “box”; Eckes, 2011). In our implementation, each person’s estimated ability was adjusted for variation in the differences in the raters’ severity as well as prompt difficulty for the AUT and TCIA.

***Abstract Painting Task.*** Participants were given five paintbrushes, an assortment of different colored acrylic paints, five paint tools, one blank 11 in. x 14 in. canvas board, and one apron. The participants were then given 15 minutes to paint freely, the only instruction being to produce an abstract painting.

Ratings were done by five human raters, who were first shown previews of all 99 paintings. They were then asked to simply rate the painting, on a scale of 1 (not at all creative) through 7 (extremely creative). In the MFRM, we specified participants and raters as facets. We estimated Rasch “fair average” scores, which were used as our overall measure of painting ability. Our five raters varied in their severity, from -0.17 to 0.67 (more lenient to more severe;



## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

in the  $Z$  metric), thus justifying the use of faceted Rasch scaling. The scores are scores after adjusting for the difficulty of the severity of the raters ranged from -1.90 to 1.68 (in the  $Z$  metric). The model's Rasch person-reliability (similar to Cronbach's alpha) was 0.78, indicating that the scoring design reliably estimated people's underlying creative painting ability. The many-facet Rasch analyses were carried out using *TAM* package (version 3.7-16; Robitzsch et al., 2021) in R (version 4.1.0; R Core Team, 2021).

***Test of Creative Imagery Abilities (Forms A and B).*** Participants all received one of two paper versions of the TCIA, a figural creativity task that asked participants to draw pictures using 7 image prompts. The prompts were combinations of 2-4 elements, either dots or lines (see Appendix A for images). Form B was simply a rotation of Form A by 180 degrees. Participants were instructed to underline the idea they liked the most and to draw and title their idea. Then, they were told that they could complete the image with unrestricted elements, as well as change and develop it to create something even more unusual.

Three raters were trained on the scoring scheme, which targeted three qualities of the finished product: vividness, originality, and transformativeness. A high level of vividness was recognized by an abundance of detail in the completion of the initial figure, a clear depiction of motion and dynamics in the drawing, or a complex presentation of metaphorical and symbolic content. A high level of originality was recognized by a depiction of new objects, activities, processes, and events in the drawing that differ considerably from the actually existing ones, surprising and novel presentation of cultural artifacts such as works of art, or amusing presentation of contents, suggesting a good sense of humor. A high level of transformativeness required multiplication (multiplying an element of the image), hyperbolization (excessive distortion of proportions, for example by emphasizing an element of the image), or amplification

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

(adding detail to the image). Participants received 0 to 2 points in each category for each drawing.

Similar to the abstract painting task, we estimated a MFRM to adjust for rater severity but also prompt severity for each TCIA rating (vividness, originality, and transformativeness). We specified participants, raters, and prompts as facets. Our three raters varied in their severity, from -0.65 to 0.49 (more lenient to more severe; in the  $Z$  metric) for the vividness rating, -0.73 to 0.89 for the originality rating, and -0.87 to 0.55 for the transformativeness rating. These varying severities justify the use of faceted Rasch scaling. The scores are generated after adjusting for the difficulty of the prompts and severity of the raters, which ranged from -2.50 to 2.33 (in the  $Z$  metric) for the vividness rating, -2.29 to 2.46 for the originality rating, and -1.93 to 2.27 for the transformativeness rating. The model's Rasch person-reliability was 0.89 for vividness, 0.84 for originality, and 0.82, indicating that there was high consistency between raters.

***Alternate Uses Task.*** To test verbal divergent thinking ability, participants were given three rounds of a computerized AUT, presented on Qualtrics survey software. They were asked to think of as many original and creative uses for objects as they could, and were encouraged to come up with responses that “strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different.” They responded to the three prompts (“box,” “rope,” and “pen”) for 2 minutes each.

Three raters assigned each response a value from 1 to 5, based on how novel and useful they were. The raters then reconciled the ratings that diverged the most.

Similar to the TCIA, we estimated a MFRM to adjust for rater severity and prompt severity. In this model, each person's estimated trait divergent thinking ability is adjusted for

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

variation in the difficulty of the prompts and differences in the raters' severity. We specified participants, raters, and prompts as facets. We estimated Rasch "fair average" scores, which were used as our overall measure of creative painting ability. Our three raters varied in their severity, from -0.90 to 0.87 (more lenient to more severe; in the *Z* metric), thus justifying the use of faceted Rasch scaling. The scores are scores after adjusting for the difficulty of the prompts and severity of the raters and ranged from -1.08 to 1.21 (in the *Z* metric). The model's Rasch person-reliability was 0.88, indicating that there was high consistency between raters.

***Fluid Intelligence Scale.*** Participants were given 3 minutes to get through as many sequence-completion questions as they could. For each of the 13 questions, three pictures are shown, and participants must select a fourth image, out of 6 options, to complete the sequence.

### ***Exploratory Measures***

***Forward Flow Task.*** Forward flow (FF) is a measure that is adjacent to verbal divergent thinking ability. It aims to measure participants' ability to connect concepts that, theoretically, are disparately related to one another in their semantic space ("semantic evolution;" Anderson et al., 2018). Our administration of the task required participants to fill 19 text boxes with words that were serially, semantically connected to the preceding word. The task began with a prompt word, of which there were three ("bear," "candle," and "table").

***Inventory of Creative Activities and Achievements.*** Another measure of creative engagement was collected by presenting participants with 65 activities across eight domains of creativity: literature, music, arts and crafts, creative cooking, sports, visual arts, performing arts, and science and engineering. For each domain, participants reported on a scale from "never" to "more than 10 times" to indicate how often they had carried out listed activities over the last 10 years. Participants also provided the number of years they estimated to have spent engaged in

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

each domain. Additionally, the participants were asked to list their five most creative achievements in their lives.

***Creative Mindset Scale.*** This scale measured whether participants' beliefs about their own creative abilities reflected a "growth mindset" or a "fixed mindset." An example of a belief that creativity is a primarily innate ability ("fixed mindset") is, "Creativity can be developed, but one either is or is not a truly creative person," whereas an example of a belief that creativity can be developed is, "Practice makes perfect—perseverance and trying hard are the best ways to develop and expand one's capabilities." These items were rated on a scale of 1 ("definitely not") to 5 ("definitely yes").

***NEO-Five Factor Personality Inventory.*** The 60-item NEO-FFI was selected to assess the five major personality factors: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Participants indicated their agreement with the items on a 5-point scale from 1 ("strongly disagree") to 5 ("strongly agree").

### ***Procedure***

Upon arrival, participants were called into the testing room and directed to a computer station, where they provided consent before continuing. The participants then completed the six measures on the computer via Qualtrics. The order of questionnaire administration was: NEO-FFI, Fluid Intelligence Scale, Inventory of Creative Activities and Accomplishments, Creative Mindset Scale, and finally all three prompts for the AUT. After completion, participants alerted the experimenter, who administered the paper version of the TCIA. Participants were then given the final 15 minutes to work freely on an abstract painting. They were then thanked and compensated for their time.

## Results

### Primary Creativity Measures of Interest

Painting scores were positively correlated with Originality ( $r = .281, p = .005$ ) and Transformativeness ( $r = .202, p = .045$ ) scores on the TCIA, but did not correlate with performance on the AUT (table 2). All three sub-scores of the TCIA were strongly intercorrelated. The AUT and TCIA only correlated with the vividness score of the TCIA ( $r = .232, p = .021$ ). Fluid intelligence was significantly positively correlated with the AUT ( $r = .293, p = .003$ ), Vividness ( $r = .322, p = .001$ ), Originality ( $r = .338, p < .001$ ), and Transformativeness ( $r = .223, p = .027$ ).

**Table 1: Pearson's Correlations**

	Painting	Vividness	Originality	Transform.	AUT	Gf
1. Painting	—					
2. Vividness	.175	—				
3. Originality	.281**	.702***	—			
4. Transformativeness	.202*	.751***	.807***	—		
5. AUT	.077	.232*	.158	.165	—	
6. Gf	.053	.322**	.338**	.223*	.293**	—

\*  $p < .05$   
 \*\*  $p < .005$   
 \*\*\*  $p < .001$

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

### **Exploratory Measures**

A full correlation matrix can be found in Appendix A. The correlation between painting ability and conscientiousness was marginally negatively significant, at  $r = -.195$  ( $p = .05$ ). Painting did not correlate significantly with any other exploratory measure.

All three subscores of the TCIA correlated positively with creative activities (originality:  $r = .307$ ,  $p = .002$ ; vividness:  $r = .260$ ,  $p = .009$ ; transformativeness:  $r = .217$ ,  $p = .031$ ) and achievements (originality:  $r = .354$ ,  $p < .001$ ; vividness:  $r = .312$ ,  $p = .002$ ; transformativeness:  $r = .242$ ,  $p = .016$ ).

Extraversion was significantly correlated with engagement with creative activities ( $r = .324$ ,  $p = .001$ ), but not achievements. Neuroticism was negatively correlated with growth mindset ( $r = -.208$ ,  $p < .05$ ). Agreeableness and openness were both positively correlated with growth mindsets (agreeableness:  $r = .315$ ,  $p = .001$ ; openness:  $r = .257$ ,  $p < .05$ ) and strongly, negatively correlated with fixed mindsets (agreeableness:  $r = -.414$ ,  $p < .001$ ; openness:  $r = -.297$ ,  $p < .005$ ). Openness also correlated with activities ( $r = .451$ ,  $p < .001$ ) and achievements ( $r = .318$ ,  $p = .001$ ).

### ***Openness to Experience Item-Level Correlations***

Looking more closely at the Openness dimension of personality, the item most widely associated with creative performance was question O3: “I am intrigued by the patterns I find in art and nature,” which was positively correlated with painting score ( $r = .254$ ,  $p = .011$ ), TCIA vividness ( $r = .221$ ,  $p = .028$ ), and TCIA originality ( $r = .280$ ,  $p = .005$ ). Question O2, “I think it's interesting to learn and develop new hobbies,” was positively correlated with TCIA vividness ( $r = .306$ ,  $p = .002$ ), TCIA originality ( $r = .224$ ,  $p = .026$ ), and TCIA transformativeness ( $r$

= .254,  $p = .011$ ). Reverse-scored O7, “I seldom notice the moods or feelings that different environments produce,” positively correlated with painting score ( $r = .235$ ,  $p = .019$ ). O9, “Sometimes when I am reading poetry or looking at a work of art, I feel a chill or wave of excitement,” correlated positively with TCIA originality ( $r = .230$ ,  $p = .022$ ).

### **Discussion**

This study demonstrates that having people produce freeform abstract paintings is a reliable and face-valid way to assess creativity. So as not to constrain participants’ painting process too much, our instructions were minimal: complete an abstract painting in fifteen minutes. All the participants followed our instructions and produced unique works of art. When three experimenters independently rated these works, two important findings emerged. First, raters, in the absence of being given specific criteria, were highly reliable in judging painting quality (“how creative do you find this piece?”). Second, the average creativity ratings given to these paintings were normally distributed. Taken together, this task seems to capture creative painting ability that varies meaningfully across the general population.

We were interested in whether two “gold standard” creativity tasks – the AUT and the TCIA – captured a domain-general ability that generalized to the painting task, or if performance on all three tasks was differentiated and domain-specific. Furthermore, we were interested in whether these tasks, which are often the sole measure of a purported domain-general creative ability, such as divergent thinking (Runco & Acar, 2012), correlate with a face-valid measure of creativity. We found some evidence that performance on the TCIA, a drawing task, more closely tracked with painting performance than the AUT, a written task, which did not correlate. However, AUT score did correlate with the “vividness” sub-score of the TCIA. If this result is

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

replicated, it might indicate that the TCIA and AUT both make use of individuals' visual imagery capacities, but in different ways.

Our exploratory findings should be interpreted cautiously since the study was not designed to test a priori hypotheses. However, our personality, mindset, and performance findings either replicate previous work or add to a small body of knowledge that investigates these correspondences. For example, Openness correlated with creative activities and achievements (Diedrich et al., 2017) and a creative growth mindset. While Openness would be expected to correlate positively with originality on the TCIA (or figural creativity, more generally) based on predictions by McCrae (1987), to our knowledge, no prior research examined the intersection of personality and figural creativity directly.

Another exploratory finding related to personality and creativity was the negative relationship between conscientiousness and painting ability. This relationship was reported in a meta-analysis by Reiter-Palmon et al. (2009), who found that splitting conscientiousness into “achievement” and “dependability” yielded a positive relationship with creative performance with the former, and a negative relationship with the latter. However, there is little research in this area.

One of the more theoretically-laden findings among the exploratory variables was the positive correlation between fluid intelligence and both laboratory measures of creativity performance. The finding that fluid intelligence correlated strongly with the AUT and all three sub-scores of the TCIA, but not with painting, adds to the mounting evidence that the former two creativity assessments target a capacity that resembles domain-general intelligence (Frith, Elbich, et al., 2021; Silvia, 2015) or may reflect a combination of shared capacities for attentional control and verbal intelligence (Frith, Kane, et al. 2021; Benedek et al., 2017). Conversely,



## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

another interpretation is that, given that the painting task correlates with the TCIA but not with other performance measures, this painting task yields a more “distilled” measure of domain-specific creativity.

Despite each measure bearing at least one relationship consistent with previous reports, there was also a notable lack of correspondence between some of these constructs. Despite correlating positively with TCIA originality and creative activities and achievements, Openness to Experience did not correlate with performance on the AUT. This finding is surprising, as this relationship is found consistently enough that it is used to infer accuracy of computerized scoring techniques relative to humans (Acar et al., 2021; Beaty & Johnson, 2021). Furthermore, painting quality did not correspond to creative activities and achievements, even when the visual arts practice subscore was isolated and compared. The absence of this relationship can be interpreted in many ways, and will itself need to be replicated before drawing definitive conclusions. However, it is possible that the quality of amateur paintings is unrelated to the amount of time individuals spend practicing their visual arts skills.

A distinct, but related, consideration is of the use of nonexpert painting raters. Although we did not set out to use the Consensual Assessment Technique to score these paintings, our protocol resembles a nonexpert execution of it. The Consensual Assessment Technique, first proposed by Teresa Amabile (1982), is a process by which creative products are rated by “appropriate observers,” or “those familiar with the domain in which the product was created or the response articulated” (p. 1001). Kaufman et al. (2011) sought to test the divergence of novice versus expert raters in poetry, and found that expert and nonexpert ratings only correlated at  $r = .71$ , with experts having higher interrater reliability. However, as paintings might be a more “accessible” artform to nonexpert raters, a similar study in the visual arts domain would be an

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

informative future direction. Furthermore, existing work has shown that those with less artistic experience produce more consistent ratings for museum-grade paintings than those with moderate levels of experience, on par with high levels of expertise (Chatterjee et al., 2010). Given that our raters achieved acceptable reliability for rating the paintings collected here, and the use of MFRMs to adjust for rater severity, the use of nonexpert raters is not necessarily cause for concern. Empirical aesthetics research, for example, often has nonexperts rate the beauty, skill, and other aesthetic qualities of paintings. Whether rating the creativity of a painting is more subjective than the beauty of a painting is an open question, our evidence nonetheless suggests that nonexperts can reach considerable consensus.

There are several ways that to improve this protocol in future work. First, providing “be creative” instructions has been shown to increase people’s divergent thinking ability on the AUT and may similarly improve people’s creativity in their abstract paintings, if such language were added to future painting instructions (Nusbaum, Silvia, & Beaty, 2014). Second, intuitively and empirically, taking longer to complete a creativity task may improve performance. Acar et al. (2021) found in their meta-analysis of 1325 verbal and 488 figural responses that longer think time predicted originality, across different divergent thinking tasks. Future studies could allot 3-5 minutes to complete the AUT and encourage participants to take the full 20 minutes to complete the seven images of the TCIA. We could also extend the time allotted to the painting task to be 30 minutes.

Given that the general “creativity ratings” were internally-consistent and varied meaningfully, an interesting avenue for future work would be to explore different rating criteria for abstract paintings and identifying separate factors or sub-scores, as in the TCIA. One possibility is to use the same scoring criteria as the TCIA. However, given that freely-painted

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

abstract art may carry more emotional associations than simple figural designs, a starting point for developing additional rating questions might consider the list of 11 “fundamental terms” to describe reactions to art that were distilled by Anjan Chatterjee and colleagues (Christensen et al., 2021; Chatterjee, 2020). Future work could determine whether any of these words could be used to assess creative aptitude or go through the process of creating a similar list of indicators of creativity with which to rate paintings. Another possibility is to use the 11 fundamental terms, many of which describe affective reactions to art, to train raters on how to judge the “expressiveness” of a piece.

Beyond further refinement of the protocol itself, and increasing sophistication of the rating schemes, there is significant room for further exploration into how creative performance on abstract freeform painting predicts performance in other creative domains and in real-world creative achievement. In the current literature that uses the AUT as a proxy for creative potential, the relationships between creative ability and creative achievement are not straightforward. For example, creative achievement may require an interplay between the Openness personality trait, general intelligence, motivation, and domain-specific expertise (Jauk et al., 2014).

In looking back to Guilford’s (1950) call to find means by which to measure creativity as an ability independent of intelligence, our abstract, freeform painting task seems to accomplish that aim, and yield reliable, normally distributed scores of creativity. Such a protocol does *not* meet the other practical requirements outlined by Barron (1965) - that many individuals simultaneously complete the task at once and be quickly scored without human raters.

As a field, though, we are not as primarily concerned with aptitude tests for the purposes of personnel selection as the early pioneers in creativity once were. The central aim today is deeper understanding of creative capacity. Painting represents a novel, yet familiar, means to

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

apprehend its nature more deeply. Surely, such a task is not as easy to implement or score as traditional laboratory-based tasks that have dominated the literature. However, based on our results, we think this protocol represents a fruitful avenue for future work in collecting and rating a sample of ecologically and face-valid creative products. Inclusion of a painting task helps illuminate the similarities and differences between performance measures, including text-based and figural divergent thinking tasks, fluid intelligence tests, and hands-on artistic activities, and offers us another path to advancing understanding.

## References

- Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C. T., & Organisciak, P. (2021). Applying Automated Originality Scoring to the Verbal Form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly*, 00169862211061874.
- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5), 997.
- Aubert, M., Lebe, R., Oktaviana, A. A., Tang, M., Burhan, B., Jusdi, A., ... & Brumm, A. (2019). Earliest hunting scene in prehistoric art. *Nature*, 576(7787), 442-445.
- Ausubel, D. P. (1967). Learning theory and classroom practice. *Ontario Institute for Studies in Education Bulletin*.
- Barron, F. (1965). Some studies of creativity at the Institute of Personality Assessment and Research. *The Creative Organization*, ed. Gary A. Steiner, 118-25.
- Beaty, R. E. (2015). The neuroscience of musical improvisation. *Neuroscience & Biobehavioral Reviews*, 51, 108-117. <https://doi.org/10.1016/j.neubiorev.2015.01.004>
- Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods*, 53(2), 757-780.
- Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic Distance And the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Creativity Research Journal*, 0(0), 1–16. <https://doi.org/10.1080/10400419.2022.2025720>
- Beaty, R. E., & Silvia, P. J. (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & Cognition*, 41(2), 255-267. <https://doi.org/10.3758/s13421-012-0258-5>

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

Benedek, M., Karstendiek, M., Ceh, S. M., Grabner, R. H., Krammer, G., Lebuda, I., ... &

Kaufman, J. C. (2021). Creativity myths: Prevalence and correlates of misconceptions on creativity. *Personality and Individual Differences, 182*, 111068.

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Thinking & Reasoning, 23*(2), 158-183.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity research journal, 17*(1), 37-50.

Chatterjee, A. 2020, July 1. *Coming to Terms with Art*. Templeton Religion Trust.

<https://templetonreligiontrust.org/explore/coming-to-terms-with-art/>

Christensen, A. P., Cardillo, E. R., & Chatterjee, A. (2021). Can art promote understanding? A review of the psychology and neuroscience of aesthetic cognitivism.

Christensen, A. P., Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2018). Clever people:

Intelligence and humor production ability. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 136.

Christensen, P. R., & Guilford, J. P. (1963). An experimental study of verbal fluency factors.

*British Journal of Statistical Psychology, 16*, 1–26. <https://doi.org/10.1111/j.2044-8317.1963.tb00195.x>

Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018).

Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts, 12*(3), 304.

Eckes, T. (2011). Introduction to many-facet Rasch measurement. *Franfurt am Main: Peter Lang*.

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

Frith, E., Elbich, D. B., Christensen, A. P., Rosenberg, M. D., Chen, Q., Kane, M. J., ... & Beaty, R. E. (2021). Intelligence and creativity share a common cognitive and neural basis.

*Journal of Experimental Psychology: General*, 150(4), 609.

Frith, E., Kane, M. J., Welhaf, M. S., Christensen, A. P., Silvia, P. J., & Beaty, R. E. (2021).

Keeping creativity under control: contributions of attention control and fluid intelligence to divergent thinking. *Creativity Research Journal*, 33(2), 138-157.

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1), 3-14.

Guilford, J. P. (1950). Creativity. *American psychologist*, 5(9).

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the “coordinates of truth”. *Perspectives on Psychological Science*, 6(3), 253-271.

Jankowska, D. M., & Karwowski, M. (2015). Measuring creative imagery abilities. *Frontiers in Psychology*, 6, 1591. <https://doi.org/10.3389/fpsyg.2015.01591>

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The road to creative achievement: A latent variable model of ability and personality predictors. *European journal of personality*, 28(1), 95-105.

Jellen, H. G., & Urban, K. K. (1989). Assessing creative potential world-wide: the first cross-cultural application of the test for creative thinking—drawing production (TCT-DP). *Gifted Education International*, 6(2), 78-86.

<https://doi.org/10.1177%2F026142948900600204>

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

Johnson, D. R., Kaufman, J. C., Baker, B., Barbot, B., Green, A., van Hell, J., ... Beaty, R.

(2021, December 1). Extracting Creativity from Narratives using Distributional Semantic Modeling. <https://doi.org/10.31234/osf.io/fmwgy>

Kimmel, A. J. (1996). *Ethical issues in behavioral research: A survey*. Blackwell Publishing.

Kriegman, S., Blackiston, D., Levin, M., & Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences*, *117*(4), 1853-1859.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology*, *52*(6), 1258.

McNemar, Q. (1946). *Diagnostic Psychological Testing; The Theory, Statistical Evaluation, and Diagnostic Application of a Battery of Tests*. Volume I.

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 231.

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 176–186. <https://doi.org/10.1037/aca0000230>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Reiter-Palmon, R., Illies, M. Y., Kobe Cross, L., Buboltz, C., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(2), 73.

Robitzsch, A., Kiefer, T., & Wu, M. (2021). TAM: Test Analysis Modules. R package version 3.7-16. <https://CRAN.R-project.org/package=TAM>



## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential.

*Creativity research journal*, 24(1), 66-75.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity research*

*journal*, 24(1), 92-96.

Silvia, P. J. (2015). Intelligence and creativity are pretty similar after all. *Educational*

*psychology review*, 27(4), 599-606.

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid

intelligence for creative thought. *Intelligence*, 40(4), 343-351.

Torrance, P. (1966). Torrance tests of creative thinking. Princeton, NJ: Personnel Press

# CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

## Appendix A: Pearson's correlations for all DVs

Variable	Painting	TCIA_O	TCIA_V	TCIA_T	AUT	Gf	Growth_MS	Fixed_MS	Open	Consc	Extra	Agree	Neurot	VisArt_Act	Activ	Achiev
1. Painting	—															
2. TCIA_O	0.281**	—														
3. TCIA_V	0.175	0.702***	—													
4. TCIA_T	0.202*	0.807***	0.751***	—												
5. AUT	0.077	0.158	0.232*	0.165	—											
6. Gf	0.053	0.338***	0.322***	0.223*	0.293**	—										
7. Growth_MS	-0.018	0.191	0.165	0.096	0.066	0.074	—									
8. Fixed_MS	-0.122	-0.168	-0.122	-0.124	-0.180	-0.080	-0.508***	—								
9. Open	0.126	0.195	0.148	0.114	0.133	-0.004	0.257*	-0.297**	—							
10. Consc	-0.195	-0.022	0.036	0.060	-0.041	-0.055	0.263*	-0.108	0.010	—						
11. Extra	-0.188	0.093	0.120	0.096	-0.039	0.099	0.181	0.074	0.210*	0.374***	—					
12. Agree	0.039	-0.127	0.019	-0.048	0.135	0.055	0.315***	-0.414***	0.090	0.177	-0.010	—				
13. Neurot	0.005	-0.076	-0.082	-0.046	-0.029	-0.151	-0.208*	0.052	0.163	-0.468***	-0.285**	-0.072	—			
14. VisArt_Act	0.112	0.208*	0.241*	0.190	0.020*	0.126	0.105	-0.082	0.245*	0.017	0.075	-0.056	0.170	—		
15. Activities	0.060	0.307**	0.260**	0.217*	0.021	0.088	0.216*	-0.096	0.451***	0.192	0.324***	0.004	0.022	0.736***	—	
16. Achievements	0.015	0.354***	0.312**	0.242*	0.130	0.151	0.209*	-0.214*	0.318***	0.037	-0.057	0.039	0.150	0.359***	0.486***	—

\* p < .05

\*\* p < .005

\*\*\* p < .001

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

### **Appendix B: Measures Used**

#### **Alternate Uses Task**

Instructions:

For this task, you'll be asked to come up with as many original and creative uses for objects as you can. The goal is to come up with \*creative ideas\*, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different.

You will be asked to type uses for 3 different objects.

You will have 2 minutes to type as many creative uses for each object as you can -- just press TAB after each one.

Click the arrow to begin.

Example Prompt:

Please list all of the creative, unusual uses for a ROPE you can think of.

Press TAB after each one.

## CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

### **Forward Flow Task**

Instructions:

In this task, starting from a given word, your job is to write down the next word that follows in your mind from the previous word. Please put down only single words, and do not use proper nouns (such as names, brands, etc.).

Press TAB after each one.

Click the arrow to begin.

Example Prompt:

Write down the next word that follows in your mind from the previous word.

Press TAB after each word. Continue when all text boxes are complete.

Your starting word is 'Table'

# CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

## Test of Creative Image Abilities

Each participant receives 7 image prompts, either A or B, each with the following instructions:

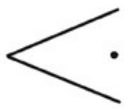




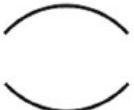




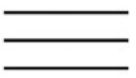

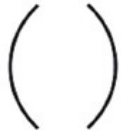

What does this drawing remind you of? Please, write down.

The more ideas, the better!

Inside the box, underline the idea that you like the most. You can complete it with unrestricted elements, change and develop it so you create something even more unusual. Please, draw your idea.

Write down the title.

Image Prompts:

TCIA A - 1	TCIA A - 2	TCIA A - 3	TCIA A - 4	TCIA A - 5	TCIA A - 6	TCIA A - 7
						
TCIA B - 1	TCIA B - 2	TCIA B - 3	TCIA B - 4	TCIA B - 5	TCIA B - 6	TCIA B - 7
						

# CORRESPONDENCES BETWEEN CREATIVITY MEASURES AND CREATIVE OUTPUT

## Fluid Intelligence Scale

Instructions:

For the following task, you will be presented with a series of drawings contained within a row of boxes. The last box in the series will be empty with dotted lines around the boarder. The row of boxes to the right of the sample are the answer choices: One of these correctly completes the series. Look at the example below. Answer choice "E" correctly completes the series.

You will have 3 minutes for this task, with a clock counting down at the bottom of the screen. After three minutes elapse, you will automatically move on to the next part. Please click the arrow to start.

Example:

