

Abstract

The *consensual assessment technique* (CAT; Amabile, 1982, 1996) represents one of the most popular evaluation techniques used by researchers to assess creative artifacts. In this paper we discuss how the prototypical use of the CAT, while useful for identifying unambiguous examples of creative artifacts, can inadvertently kill the curiosity of researchers interested in interpreting and understanding more ambiguous, contested, and divergent examples of creative expression. More specifically, we open the article by briefly describing the CAT and how it can simultaneously help and hinder researchers in making judgments about the creativity of artifacts in and across domains. We then introduce a methodological elaboration on the CAT, called the Divergent, Open-Ended and Generative (DOG) approach to interpreting creative artifacts. Next, we present a secondary analysis of CAT data to illustrate the DOG approach and how it can extend insights about creativity when used in conjunction with the CAT. We close with a discussion of implications for creativity theory and research.

KEY WORDS: *Consensual Assessment Technique, Creativity Assessment, Methodology, Ambiguity, Secondary Analysis of Data*

Is Curiosity Killed by the CAT?

A Divergent, Open-Ended and Generative (DOG) Approach to Creativity Assessment

Creativity assessments can be thought of as sensemaking tools that researchers develop and use to inform their judgments about the creativity of ideas, actions, or outcomes. Although creativity assessments are helpful in making judgments about creative outcomes, how creativity researchers use those assessments have nontrivial consequences for how creative phenomena are represented and understood (Barbot et al., 2019; Long et al., 2022). Consequently, creativity researchers have, in recent years, stressed the importance of taking steps to ensure that creativity assessments produce more reliable, valid, replicable results (Makel & Plucker, 2014). Issues of reliability, validity, reproducibility of research results based on creativity assessments are clearly important. That said, when creativity researchers focus so much of their attention on ensuring that their judgements of creativity are based on unambiguous assessment results, they may inadvertently overlook important insights about creativity that can be found in more ambiguous and contested results of those assessments. Indeed, viewing ambiguous or contested results as “statistical noise” can obscure important aspects of creative expression and unnecessarily limit our understanding of what counts as creative. Moreover, as we will discuss, doing so is somewhat antithetical to our conceptual and theoretical understanding of creativity itself.

The aim of this paper is to discuss how taking a broader, “both/and approach” to interpreting the results of creativity assessments can advance the fields’ understanding of creative phenomena. In this paper we focus on the assessment of creative artifacts and more specifically the *consensual assessment technique* (CAT; Amabile, 1982, 1996). We zoom in on creative artifacts because they represent the socio-material outcomes of creative efforts and thereby serve as the central “objects” of sensemaking interpretation when making judgments

about creativity. We also focus on the CAT because it represents one of the most popular assessment techniques used by researchers to assess creative artifacts—what some creativity researchers have described as the closest thing to a “gold standard” that the field has for assessing creative products (Baer & McKool, 2009; Carson, 2006).

We open the article by briefly describing the CAT and how it can simultaneously help and hinder researchers in making judgments about the creativity of artifacts in and across domains. We then introduce a methodological elaboration on the CAT, called the Divergent, Open-Ended and Generative (DOG) approach to interpreting creative artifacts. Next, we present a secondary analysis of CAT data to illustrate the DOG approach and how it can extend insights about creativity when used in conjunction with the CAT. We close with a discussion of implications for creativity theory and research.

How the CAT Kills Curiosity

The CAT involves impaneling a group of relevant experts who are tasked with independently rating the relative creativity of a set of artifacts using their own implicit understanding of creativity. The validity of the ratings is warranted by the expertise and consensus of the raters. Researchers determine the degree of consensus by calculating inter-rater reliabilities amongst the judges (Amabile, 1982; 1996; Hennessey, 1994), which tends to be high amongst expert raters (Kaufman et al. 2008). Given the rather elegant methodical logic of the CAT, it is not surprising that it has become one of the most popular methods of assessing the creativity of artifacts.

In fact, the CAT is widely used in the creativity literature helping the evaluation of a variety of artifacts from various domains and cultures (e.g., Kaufman et al., 2009; Hickey, 2001; Hennessey et al., 2008). This is not to say that the CAT is without its critics. Indeed, as with

any methodology, there are limitations and variations in how it is used. Researchers have thereby endeavored to improve the reliability and validity of judgments based on its use (see Cseh & Jeffries, 2019). Although such efforts can improve the psychometric properties of results derived from the CAT, we see a more fundamental issue with respect to the CAT (and related assessment techniques), namely, its focus on *consensus* and convergence.

Given that creativity is associated with novelty and originality (Acar et al., 2017), creative artifacts and their interpretation are often contested. In fact, we assert that it is precisely the diversity of interpretations or perspectives on specific artifacts that enhances their creative value. Researchers tend to recognize the necessity of originality in their conceptualizations of what makes a product or artifact creative (see Runco, 1991; Runco & Acar, 2012); however, the same cannot be said about how the CAT ultimately treats the judgments of raters who evaluate creative artifacts. Although the CAT relies on initial, implicit judgments of creativity amongst different raters, the methodology then focuses on consensus of judgments when it comes to ultimately determining what artifacts in a sample are determined to be creative. Agreement on a creative artifact is paramount because that agreement results in the highest mean creative rating for a given product. Any deviation from that agreement by a rater will result in a creative product to a lower mean rating. In this way, the CAT privileges unambiguously creative outcomes (i.e., those artifacts that have achieved high levels of statistical agreement amongst raters). This emphasis on consensus, somewhat ironically, stresses the importance of the sameness of perceptions amongst judges and essentially treats all other artifacts in a sample, which don't reach a consensus of creativity, as not creative or not as creative as those that achieved a high level of agreement. This approach, when taken at face value, makes sense with respect to the psychometric imperative of ensuring that scores from the CAT are reliable. Those reliable scores

are then used to pursue a variety of important research questions and relevant statistical models to learn about the creative process, the individual or collaborative team, and the environment.

Why should we view the CAT's emphasis on the importance of attaining high levels of consensus among raters as problematic? If we start with the premise that differences of perspectives about what is created is a key marker of creative works, then we can recognize that differences of perspective are important for the creative process itself (Author 3, 2020; Author 3, 2019). This is because there are always various points of view amongst different raters regarding the qualities of creative products. At this point, an example may help further clarify why we view the emphasis on consensus in the CAT methodology as problematic.

What Divergent Perspectives Can Reveal

Consider case-studies (Author 3, 2010; 2012) that report on the assessment of creativity in Easter egg decorations wherein both similarities and differences could be observed between the assessment of ethnographers, priests, artists, and craftspeople. More specifically, the positive evaluation of creativity in this craft depended primarily on whether the person doing the assessment was involved in decoration or other artistic activities. The 'view from inside', specific for artists and craftsmen, is marked by an appreciation of creativity by highlighting its aesthetic qualities and the risks involved in decorating such a small and fragile object. On the other hand, the 'view from outside', specific for priests and ethnographers, emphasizes other criteria, namely the role of tradition and the small (and generally unwelcomed) departures from it. What the case-studies demonstrate is that creativity evaluations are rooted in the knowledge, expectations, experience, and values of the individuals formulating them and, importantly, of their wider communities (in this case, professional groups).

However, if we were to consider the evaluation of Easter egg creativity with the help of a CAT, the outcome would have been different. First, the CAT tends to discourage the selection and use of evaluators coming from different groups and communities that have various ways of relating to the creative artifact. Homogeneity of evaluators is imposed in the CAT methodology by restricting the raters to people considered ‘experts’ – those with general expertise in the area of the creative product, whether formal or informal – and ignoring the fact that experts themselves belong to different communities and that, in everyday life, it is not only experts who evaluate creativity and contribute to the celebration of creative ideas and artifacts.

Second, while products are evaluated or ranked against each other in the CAT methodology, ensuring that evaluations remain contextual, the statistical procedures used in the CAT aim towards convergence – hence the name of the method itself – meaning that discrepancies between evaluators, if they exist, are statistically ironed-out to reach a consensual set of scores. If the discrepancies between evaluators lack acceptable levels of average agreement, then the process might be repeated to reduce the ‘statistical noise’ in the data rather than view these differences as valuable sources of information for creativity researchers. We are not arguing that the prototypical use of the CAT lacks value or meaning. Indeed, it is quite useful, whenever we want to understand how much experts agree on what is unambiguously creative and, thus, obtain a final score or rank for our creative artifact. We argue that this approach goes against exactly what makes creative artifacts ‘creative’ which is the discussion and debate they generate. In other words, this method ignores the divergence of views about creativity and its importance for the way in which the artifact is received, used, and transformed.

When Artifacts are Ambiguously Creative

At best, the CAT focuses us on discovering what most people agree that is creative without looking deeper into reasons why something is considered this way, which might differ, and especially devaluing artifacts that are seemingly more controversial or ambiguously creative. Given the conceptual and empirical importance of tolerance for ambiguity and uncertainty in the creative process (Author 2, in press; Furnham & Marks, 2013; Zenasni et al., 2008), we argue ambiguity should be embraced in the evaluation and investigation of creative artifacts. The capacity to embrace ambiguity should be paramount for CAT raters of creativity and creativity scholars using the CAT to research creativity in specific contexts. In everyday contexts, we argue, artifacts that are more controversial or ambiguous are in the minority and inflated by the statistical and methodological techniques we use to boost consensus at the expense of divergence. This is paradoxical, given the role of divergence for creativity as a whole as we explained earlier.

By acknowledging the role of divergence in creativity only when it comes to creative action and downplaying this role when it comes to evaluating creativity, we are reinforcing the view that there is little to no creativity in the actual interpretation of a creative artifact. This is demonstrably problematic, as we can see from engaging with other literatures from the dynamic definition of creativity and its emphasis of open-ended estimations (Corazza, 2016) to extensive work on user innovation (von Hippel, 1988) and the collaborative, distributed, and often interdisciplinary nature of how creative ideas and accomplishments evolve across people and time (Clapp, 2016; Ness & Søreide, 2014).

Put simply, the prototypical use of the CAT can kill the curiosity of researchers interested in understanding the more nuanced, contested, and divergent perceptions of creativity amongst

raters. How then might creativity researchers use the results from the CAT to understand the more ambiguous and contested forms of creative expression?

Introducing the DOG Elaboration

As mentioned, our aim in this paper is to challenge the largely unquestioned emphasis on consensus in creativity assessment and, in doing so, complexify our understanding of creativity evaluations as a creative or generative process. To be sure, we are not arguing for abandoning the CAT; rather, we realize that it serves specific purposes and comes with a constraining set of assumptions that are overdue for inspection. Moreover, the methodological process the CAT entails offers uninspected statistics and opportunities for multi-dimensional and multi-method extensions to further our understanding about creative artifacts and the people, process, experiences, and environments that shape them. Most importantly, the CAT eliminates a researcher's curiosity about why people evaluate creativity the way they do, why they might disagree about these evaluations, and what happens to artifacts that are ambiguously creative. For this purpose, we advance the DOG—the Divergent, Open-Ended, and Generative interpretations of creative artifacts. Importantly, the DOG is an enhancement, not a replacement of the CAT. The DOG represents a secondary analysis technique that allows researchers to identify and interpret divergent ratings of artifacts.

Specifically, the DOG proceeds through a series of steps outlined below, which researchers conduct after a group of three or more raters have scored an entire sample of creative artifacts using the CAT. The sorting process incorporates the mean ratings across raters and the variance (standard deviation) of those ratings across raters. More raters should result in greater potential for divergent perspectives and a larger distribution of variance among rated artifacts; we recommend using at least three raters, and as many as five to seven, because it is in alignment

with the prototypical CAT procedure and offers the possibility of more divergent perspectives, which is important both for identifying high levels of consensus and for identifying divergence amongst raters. Five to seven raters is not *necessarily* optimal (as this is an open empirical question at this point). We therefore recognize that it may be possible to effectively employ the DOG approach with fewer or greater numbers of raters.

The first step in the DOG methodology is to recognize that artifacts that have been rated using the CAT can be classified into at least five different conceptual categories based on a combination of the average CAT rating (relatively high creativity rating versus relatively low creativity rating) and the level of agreement amongst CAT raters (relatively high agreement versus relatively low agreement). As illustrated in Figure 1, we specified five categories for this initial investigation: *unambiguously creative* (i.e., high creativity rating, high consensus amongst raters), *ambiguously creative* (i.e., high creativity rating, low consensus amongst raters), *unambiguously uncreative* (i.e., low creativity rating, high consensus amongst raters), *ambiguously uncreative* (i.e., low creativity rating, low consensus amongst raters), and *ambiguous potential creativity* (i.e., moderate creativity rating, least consensus amongst raters). We recognized that other categories can be specified and examined, such as average ambiguously creative, but for the purpose of introducing the DOG approach we focused on the five categories depicted in Figure 1. In what follows, we briefly describe how we used existing CAT scores and variability amongst raters to classify artifacts into each of the five categories (see also Table 1).

Unambiguously creative (UC) reflects high ratings of creativity (e.g., top 10th percentile in creativity scores) with high levels of consensus among raters (e.g., $SD < 0.25$)¹. This category

¹ The use of the 10th percentile and $SD > 0.25$ are suggested here as illustrative examples for how researchers could set the cutoff. Researchers' decisions will need to consider the sample size and the response scale. The size of the

reflects consensus amongst raters as to the most creative artifacts within the sample. Although we use the top 10th percentile and an a priori standard deviation threshold to establish this category, we recognize that the actual percentile selected for this step is at the discretion of researchers and should be based on the size of the sample and rating scale in order to generate meaningful analytic subsamples of artifacts in each category and to conduct the secondary analyses of interest (we propose that subsamples of 10 artifacts offers a rich set for secondary analysis). Within that top 10th percentile subsample, researchers can then rank each rated artifact by standard deviation across raters and either use an a priori cutoff or adjust depending on the analytic sample that results. For instance, the unambiguously creative artifacts can be selected with a cutoff (e.g., $SD < 0.25$) or simply identified as the top 10 artifacts with the lowest SD. This reflects an important distinction offered by the DOG methodology (compared to the CAT) as it differentiates the subsample of artifacts that raters agree represent the highest levels of creativity in the sample.

Along similar lines, *ambiguously creative* (AC) artifacts reflect products rated as both highly creative *and* highly contested because the mean rating across raters reached an a priori cutoff of high variability (e.g., $SD > 1.0$) among the top 10th percentile of mean creativity ratings. Again, this is a distinction typically not made using the CAT methodology and reflects a novel contribution of the DOG because it highlights potentially important differences in views amongst raters.

SDs for creative artifact scoring depends on the number of points used on the rating scale. If an artifact's ratings have an SD of 1, this might translate to an inter-rater agreement of .50 for that artifact when using a 5-point scale. If using a 7-point scale and the SD is 1, the inter-rater agreement could increase to .75, which may not be a meaningful threshold to capture ambiguity. Researchers could consider using a 10th percentile cutoff for both standard deviation and mean to determine their analytic sample.

At the other end of the creativity rating distribution the next two categories—*unambiguously uncreative* (UU) and *ambiguously uncreative* (AU)—are identified in an identical process as the first two categories; the difference is these categories are established using the artifacts in the bottom 10th percentile of mean creativity ratings. UU artifacts have low mean creativity ratings and high levels of consensus amongst raters (e.g., $SD < 0.25$), whereas AU has low mean creativity ratings and low levels of consensus amongst raters ($SD > 1.0$).

The fifth, and final, category we propose—*ambiguous potential creativity* (APC)—represents those artifacts with the highest level of difference among raters based on the largest standard deviations. Due to these high standard deviations, these artifacts will likely fall between the 25th and 75th percentile in mean creativity rating. From the perspective of the DOG, the APC category represents one of the most compelling categories because it reflects the highest levels of difference both in ratings of creativity and in agreement amongst raters. Table 1 provides the definition and operationalization of these DOG categories.

Once these subsamples of artifacts have been established, secondary DOG analysis can be conducted using different quantitative and qualitative methodologies, depending on the questions of theoretical and empirical interest. In this way, the DOG opens a broad horizon of methodological possibilities and insights. For instance, remaining within the quantitative methodological tradition of the CAT, if the subsamples offer enough analytic power, researchers can make comparisons about person-level characteristics of the creators. Researchers might ask if certain creative self-beliefs (Author 2, 2018), such as creative confidence beliefs (Author 1, 2021), creative metacognition (Author 2, 2013; Author 1, 2021), intellectual risk-taking (Author 2, in press), or creative anxiety (Daker et al., 2019), differ in a meaningful way for the creators producing UC, AC, AU, UU, or APC artifacts. Researchers could also ask questions about the

creative process to generate artifacts in the different categories, such as the amount of time or effort committed.

The DOG establishes an analytic framework that can extend to a variety of methodologies within the qualitative tradition. For instance, from a sociocultural perspective (Author 3, 2020), researchers could analyze how the cultural background characteristics and experiences of creators in each category differ and what role those characteristics may play in the ambiguity of the artifacts they create. Researchers could survey or interview both the creators and the raters to learn about the role of sociocultural perspectives and context. Through a phenomenological and emergentist lens (see Author 1, 2018), researchers could contrast the meaning-making process of creators across the categories through the creators' post-production reflection on the creative process. Did creators' affective experience differ for more ambiguously creative artifacts? Did the development of a narrative in the creative work evolve differently? Did the confidence in their creative work differ for creators of UC and AC artifacts? To what degree was the ambiguity in creative artifacts intentional, accidental, or due to a lack of time to finish? Is there any evidence of a potential sociocultural mismatch between the expert raters and the creators (e.g., ethnic, linguistic, cultural, or generational)?

Moreover, the categories open new windows of insight and analysis as to assessments of creativity, including identifying potentially interesting and important cultural and contextual insights and perspectives about what constitutes a creative artifact. Researchers can, for instance, use one or more of the ambiguously creative DOG categories in their tertiary analysis by having people in and across different contexts and from different vantage points examine and explain what they view as creative or not creative about these artifacts. This can include

returning to the original pool of expert raters and having the raters discuss the rationale and difference amongst their original ratings.

Some of the possible follow-up questions posed by the DOG may simply require additional planning in the research design prior to conducting the CAT, such as additional surveying or interviewing of both the raters and creators. With care and attention to research procedures and documentation, we believe the DOG can open the door and welcome a variety of methodological traditions and epistemological perspectives to the study of the phenomenon of creativity to advance the understanding and relevance of the field of creativity research.

The DOG in Action: A Secondary Analysis of CAT Data

In the present study, we unleash the DOG methodology on an extant dataset of CAT creativity ratings used in previously published studies (Author 1, 2021a; Author 1, 2021b). The creative artifacts are the work of individual U.S. middle school students developing their own imaginary creature in both illustration and written description. This study is meant to be illustrative of the DOG approach's potential by demonstrating the kinds of insights and questions that can be derived from re-analyzing a subsample of student creators of artifacts in each DOG category. Our secondary analysis also includes using creative metacognition ratings, creative self-efficacy ratings, and self-evaluation of the creative artifacts. To illustrate the application of qualitative analyses, we also analyze artifacts among the different categories to highlight sources of ambiguity and complexity at work in initial ratings, how these sources of ambiguity and complexity can be revealed by the DOG, and what that ambiguity may reveal about the phenomenon of creativity, more generally.

Method of the Original CAT Study

As reported in the original study (Author, 2021), the results of the initial CAT analysis in our extant dataset demonstrated strong intraclass coefficient (ICC) reliability. More specifically, using a two-way random model (Shrout & Fleiss, 1979) and interpretation guidelines offered by Landis and Koch (1977), the ICC for assessing the (a) creative illustration was $ICC = 0.88$ for Group 1 raters and $ICC = 0.87$ for Group 2 raters; (b) creative description was $ICC = 0.92$ for Group 1 raters and $ICC = 0.90$ for Group 2 raters; and (c) combined creativity rating was $ICC = 0.91$ for Group 1 and $ICC = 0.90$ Group 2 raters.² Those original findings produced meaningful results from the CAT. First, they provided multiple sources of convergent validation to the creativity scores and a high degree of inter-rater reliability. Second, they illustrated the mediating role of creative self-beliefs and creative metacognition to stronger creative production. Therefore, we believe secondary analysis of the original study presents an optimal opportunity to employ the DOG approach given the rigorous and successful application of the CAT.

Sample

The original participant sample (Author 1, 2021) included $N = 872$ students in Grade 8 attending eight middle schools in the Pacific Northwest of the U.S. The sample was diverse for the region, with 65.9% of students identified as white, 12.6% receiving special education services, and 54.4% of students marginalized economically. Of the non-white students, 21.3% were identified as Hispanic, 8.1% multiracial, and between 0.3–1.6% Pacific Islander/Native Hawaiian, Black, Asian, or Alaska Native/American Indian.

Procedures

Students completed the full assessment protocol in a 45-minute normal class period on their own using pencil and paper. Teachers read aloud specific instructions to ensure that

² For full results of validity and reliability of these creativity ratings, we refer readers to Author (2021). We will identify this study after the double-blind review is complete.

students were prompted in a consistent way across classrooms and schools. Instructions asked students to respond to five creative self-efficacy questions, and then invent a mythological creature and draw it without adding color. After drawing their creature, they were instructed to describe the creature in writing, then reflect on the experience of creating their creature, thinking about what worked well and what they might change next time. Next, the protocol asked students to rate the creativity of their own creature on a scale from 1–10. To manage the workload, two teams of three raters scored half of student artifacts (i.e., ~425 responses for each rater). This process ensured that raters were blind to the school origin of each response. To categorize creative metacognition, efficiently, one rater per team identified the category or categories that each student reflection fit best, including: (a) self-awareness of strengths, (b) self-regulatory, (c) contextual knowledge, and (d) creative strategy selection. Those categories were derived from the guiding CSB framework (Author 2, 2017). The raters judged the level and depth of evidence of their creative metacognition on a scale from 0–2, with zero representing no response, 1 representing partial evidence, and 2 representing full evidence.

Measures

Creative Self-Efficacy

In the original study (Author 1, 2021), the protocol involved asking students to rate their CSE prior to beginning the creative task (as recommended by Bandura, 2012 and Author 2, 2018). The authors of the original study also followed Bandura's (2005) recommendations to construct a CSE scale that would align to the task and be developmentally appropriate for adolescents. Five creative self-efficacy items were used that included a broad scale (0-10) for students to self-report their confidence to invent a unique mythological creature. Five questions

asked students to rate their confidence to invent a creature, think of details, complete a drawing, and produce something unique and creative.

Imaginary Creature Task and CAT Creativity Rating

The creative task in the initial study included both illustration and written components to ensure accessibility for students with diverse abilities in drawing and writing and to offer a more complex look at their creative production. The content of the task (i.e., invent a mythological creature) was designed to be open-ended and imaginary—similar to tasks in past research (e.g., Ward, 1994)—and relevant and accessible to middle school students, offering broad cultural and personal interpretation. The researchers in the initial study adhered to guidelines for the CAT (Amabile, 1982; Baer & McKool, 2016). The CAT technique recommends that raters have some level of expertise in the domain of interest (Amabile, 1982; Kaufman, Baer, Cole, & Sexton, 2008). Raters were recruited from teaching artists with experience integrating the arts into middle school teaching and learning. We assumed an adequate level of expertise in the domain of K-12 arts education and youth creative development because all raters had at least five years of experience producing artistic works in visual, musical, and/or theatrical domains and at least five years of experience or formal training in teaching artistic practices to K-12 students. The task prompted students to produce an illustrated and written response to ensure that detecting students' creative ideas did not rely entirely on drawing or writing skills alone.

Creative Metacognition Rating

Immediately after completing the creative task, the assessment protocol drew on students' CMC to record reflections about what worked well, why it worked, and what could be improved. Specifically, students responded to two broad reflection questions targeting creative metacognition: (a) to think about their approach and describe what worked well and (b) to think

about what they would do differently to make their creature more creative. These broad questions assessed students' self-awareness, self-regulation, contextual knowledge, and strategy selection. Creative metacognition of student reflections was scored following the same procedures and calibration process on a scale from 0–2, with 0 meaning incomplete response, 1 meaning a partial response, and 2 meaning a complete response. If a student responded to the reflection prompts, raters evaluated if it reflected at least one of the following categories (Author 2, 2017): creative strategy selection, self-regulatory monitoring, self-awareness of strengths or weaknesses, and contextual knowledge about the challenge. The difference between a rating of 1 and 2 indicated the strength and depth of the response as well as the inclusion of multiple CMC categories.

Analytic Approach of the DOG Secondary Analysis

To accomplish the first aims of producing a thorough description of the student creators who made up each DOG category, we produced a comparative table of descriptive statistics for each group of 10 students whose artifacts had a mean and SD of creativity ratings that fell into the criteria for that category. We reported the results of that visual comparison to initially evaluate if any patterns of interest could inform the qualitative portion of the DOG procedure. Next, we analyzed the artifacts within the ambiguous groups to identify sources of ambiguity and complexity, referring to both the description and illustrations produced by student creators and rated by the two panels of experts. This approach built on the long legacy of using illustrated artifacts from students to respond to research questions relevant to theory and practice (see Ganesh, 2011). These patterns are compared across the different groups of artifacts.

Results of DOG Secondary Analysis

The results below follow a sequential mixed method approach, using quantitative analyses to inform the qualitative analyses. Using the DOG approach in this context, we emphasize the qualitative analysis above the quantitative analysis but integrate the findings from both to enrich interpretation and understanding (Johnson & Onwuegbuzie, 2004; Creswell & Plano Clark, 2017). We are focusing on (a) the emergent themes that arise from this specific data set and (b) how those themes may inform the use and insights drawn from the DOG technique in other study contexts (Table 3).

DOG Subsample Composition and Comparisons

After completing a descriptive table with the person-level characteristics that compose each subsample (Table 2), we looked for patterns that might reveal questions to pursue in the more in-depth qualitative stage of the DOG approach. We included potentially meaningful characteristics of the subsample composition, including: The (a) mean and standard deviations of creativity ratings, (b) the mean of the standard deviations across the three ratings for each artifact, (c) the percent of students who were female, (d) the percent of non-white students, (e) the percent of students facing economic disadvantage, (f) the percent of students in special education, (g) the percent of English Language Learner students, (h) the mean creative self-efficacy self-rating, (i) the creative metacognition rating (by external raters), and (j) the self-ratings of creativity students gave their own mythological creature (by each student).

Variance in Ratings

The ambiguously creative ($SD = 1.15$), potential creativity ($SD = 1.69$), and uncreative ($SD = 0.87$) subsamples each had high variance in ratings. Although the variance was mostly distributed across raters, we did identify cases where the variance was due to one rater within a subsample.

Difference in Student Subsample Characteristics

The subsample composition of the UC and AC groups were very similar on most characteristics. The components that seemed to differ were in self-identified demographic variables. The unambiguously creative artifacts were completed by a subsample that was 90% female and 70% white; whereas the subsample that completed the ambiguously creative artifacts was a 50-50 split of female and male, which was also 100% white. Though the creative self-efficacy levels were nearly identical, the mean self-evaluations of creativity of the subsample who made the unambiguously creative artifacts was close to a full standard deviation higher than the self-evaluation of the ambiguously creative artifacts. The externally rated mean creative metacognition evident in students' reflections was also nearly a standard deviation higher for the unambiguously creative subsample than the ambiguously creative sample. Only one student in the ambiguously creative group rated their artifact lower than a level of 6.

The subsample producing the ambiguous potential creativity (APC) artifacts were different than any other subsample on two characteristics—being racially and ethnically diverse (mostly non-white) and having the highest proportion of students facing economic disadvantage and hardship. The average standard deviation of this group of artifacts was very high across ratings of creativity. Disagreement was often with only one rater, downgrading the rating, but sometimes included disagreement across all three raters. Interestingly, the mean creative self-efficacy for creating an imaginary creature was practically identical to the UC and AC subsamples. Their mean creative metacognition level was nearly identical to the level for the AC subsample, and the self-evaluation of the creativity of their artifact was very similar to the level of the AC subsample.

The subsamples producing the ambiguously uncreative (AU) and the unambiguously uncreative (UU) artifacts were similar on some components and different on others. The groups producing AU and UU artifacts had higher proportions of male students, higher proportions of economically disadvantaged students, and the AU group had the highest proportion of special education students of any group. The UU group was entirely white. The creative self-efficacy, creative metacognition, and self-evaluation of the creativity of artifacts seemed to trend higher for the AU group; however, the standard deviations were very high on these factors for both subsamples. That high degree of variance for the AU subsample was due to three students rating their own artifact at a 10 out of 10 on creativity, reflecting a very different assessment than the evaluation of the external raters.³ Similarly, in the UU subsample several students gave their own artifacts high ratings, such as an 8 and 10 out of 10.

Sources of Ambiguity and Complexity in Creative Artifacts

Table 3 describes different sources of ambiguity that emerged from the analyses detailed in the following section using more general descriptions to serve as an analytic guide for future research. Sources of ambiguity in the ambiguous creativity (AC) and ambiguous potential creativity (APC) categories provided an array of possibilities from incomplete and open-ended compositions to written descriptions that are generative leaving the reader and audience with questions to ponder and wonder about. For instance, one artifact (Artifact AC #1, see Figure 2) offered less detail in a somewhat incomplete composition about a creature the student has named “Keneis.” The student inserted ambiguity into the creature’s story at the beginning with “...we don’t know where she came from.” The student leaves many illustrated features, such as the

³ Due to issues in archiving the original drawings of students, some of the artifacts selected in the DOG categories were not available to for secondary analysis. The issues may have resulted from the original artifacts not being scanned appropriately or labeled with verifiable identification in the secondary dataset.

creature's face, undetailed, which could be evaluated as incomplete. However, that ambiguity in the drawing is made generative in the description by details, such as "she always has a poker face, and just kinda doesn't care" and "she doesn't seem to die." The details in the illustration are different than what the audience finds in the written description. It is possible that the divergence of what the audience may come to expect in the written description based on illustrated elements comes with too much surprise or incongruence. The element of divergence and the unexpected may reach an uncertainty threshold for some audience members and result in reduced rating of creativity.

Novelty and Imaginativeness

Another source of ambiguity and complexity may have been highly novel and imaginative elements, such as made-up words, which can be a complex act of creativity. For example, in Artifact AC #2 (see Figure 2), the student creator made up two words that appear to serve as a taxonomy for the species and characteristics of the creature ("...you can tell it's a leogeonious and not a dongleonious because of the horn on its head.") In this example, the mythological creature has a variety of details but little else is drawn around the creature; however, the student generates complexity in the narrative around the creature by the addition of this proposed taxonomy.

Humor, Cleverness, and Contradiction

Humor and cleverness in students' creative work may also be a source of ambiguity and complexity, and, in some circumstances, draw out biases from a rater within the CAT process. For instance, in Artifact AC #3 (see Figure 2), the creator makes the mythological creature a flying ice cream cone—a novel idea within the sample. The written description cleverly lists the weaknesses as "...people who eat it, chocolate syrup, and warm places like Florida." It is

possible that one rater felt this artifact was too simple and did not value the cleverness as highly as other raters as a matter of aesthetic taste. In a similar way, the complexity of some creatures and drawings may cause a rater to perceive the creation as lacking coherence. For instance, in Artifact AC #4 (see Figure 2) the student blends body elements of a giraffe, snake, and bird and adds a tattoo of “Family” to the hand of the creature, writing “...it says this because that’s what matters most to him...” He names the creature “Girdisnake.” Yet another example, Artifact AU #1 (see Figure 3), the student intentionally did not include an image and explained in the caption, “It’s a ghost so you can’t see it.” That students’ approach could be taken as completely inappropriate for the task by one rater and completely novel and effective by another. Taken together, these elements may be perceived as incoherent to some viewers or too open-ended. To others, it may have represented cleverness.

A similar source of ambiguity to humor may be contradiction in elements of an artifact. For instance, in Artifact APC #2 (see Figure 3), the creator made a flying snowman that breathed fire, which, though creative, could be too contradictory for a viewer to give a high rating. A third example of humor’s role in ambiguity of rating creativity is Artifact APC #3 (see Figure 3). The students’ description is somewhat nonsensical (“...they have a dog named cat...”) but the creature may have been funny-looking to some raters, and it is clear the creator took care with several visual elements to achieve that effect (e.g., differently-sized eyes, squiggly lines on tongue and stomach, and slightly upturned and downturned mouth). To achieve a humorous effect, the creator of Artifact APC #3 may have purposefully provided minimal detail about their creature named “Bloob,” which could be interpreted by some raters as incomplete.

Unusualness and Abstractness

Similar to humor, unusualness seemed to be a source of ambiguity in the ratings for the AC artifacts. As with the element of humor, unusualness or abstractness may have been an underlying issue for Artifact AC #5 (see Figure 3), which depicts a “potato-snail thing” that flies in outer space. The creature is unusual, comical, and whimsical, which may not fit the expectations of a viewer when evaluating and comparing the creativity of mythological creatures. It is possible that a viewer unconsciously judges the level of creativity on adult expectations that misalign with some adolescent expectations for creativity. In a more abstract example, Artifact AC #6 (see Figure 4) is titled “an eye squid king” who “rules with intelligent thoughts and words.” The illustrated creature is challenging and abstract with squiggly legs dangling from a circular body with an eye that takes up most of the bodily form. There are also small fins, a squid beak, and a crown on top. The description adds a little more context, but choices in the description are also unusual (e.g., describing the kinds of thoughts and words the creature uses).

Divergent Interpretations of the Prompt

Another source of ambiguity in judging the creativity of the creature seemed to arise when the idea of what a creature could be was very divergent from the typical response. For Artifact APC #1 (see Figure 4), the student imagined their creature to be a “book called Smile.” While two raters rated this artifact 4 out of 5, one rater gave a rating of 1 out of 5, illustrating that the unusualness of the idea and artifact may have made that rater think it was an inappropriate response. It may not come as a surprise that the ambiguous potential creativity group of artifacts was produced by the most racially and ethnically diverse group of student creators. For instance, Artifact APC #1 was produced by a Pacific Islander/Native Hawaiian

student. It is possible that different interpretations of the prompt are the result of different cultural meaning attached to key words in the prompt, such as “creature.”

Discussion

This study sought to understand what can be discovered about the divergent, open-ended, and generative dimensions of creativity that arise in the disagreement and contradictions between raters of creative artifacts. We argue these fundamental elements of creativity are lost in the CAT process. Using an extant dataset of CAT ratings of creative artifacts of middle school students, we analyzed the responses that demonstrated the greatest variance in CAT creativity ratings to illustrate what can be found in this liminal space of difference. Our aim was not to provide a definitive, alternative interpretation of artifacts but rather highlight how using the DOG methodology opens-up new possibilities for researchers to explore, identify, and interpret creative artifacts that might otherwise be lost when using the CAT alone.

This first demonstration of the DOG raises several questions about the CAT that are important to discuss. Beyond our initial assertion that the CAT privileges some aspects of creativity, namely, consensus amongst raters (as opposed to difference), we have demonstrated how the DOG can help researchers identify features of creative artifacts that are associated with consensus, such as: completeness, elaboration, and refinement. The DOG also highlights how more ambiguous and potentially important features of creative artifacts, such as humor, surprise, complexity, and divergence can be minimized and overlooked when relying solely on the CAT. These divergent, ambiguous, and surprising features have been recognized as important to creativity (for an early discussion see Bruner, 1962), however, our analysis suggests that relying only on the CAT likely obscures these potentially important creative attributes. We draw on our

findings to discuss some of these issues in the sections that follow. We then close with implications and directions for future research.

Does that CAT Privilege Completeness (and Non-ambiguity) in Creativity?

Providing a complete illustration and narrative about the creature appeared to be highly valued in the way the reviewers rated the artifacts and was a point of agreement amongst raters. Relatedly, the UC group, whose creatures were scored the highest in creativity with agreement across raters, demonstrated the highest levels in self-evaluation of the creativity of their product and in their creative metacognition regarding their awareness of the context, strategies, and personal strengths to tap for this creative challenge—they knew how to score high and that they hit the mark (Author 1, 2021). Completeness may have been strategic for students, especially given the school context in which the task was performed, where usually, incomplete work is penalized. Completeness of the artifacts for the reviewers, however, may relate to their tolerance (or intolerance) for ambiguity (Furnham & Marks, 2013), where the ambiguity inside some of the creative artifacts may have crossed an uncertainty threshold for reviewers. It is possible that the evaluative process of judging the creativity of products may tap into individuals' need for closure, resulting in a heavy reliance on strong stereotypes about creativity when more ambiguous and challenging artifacts were confronted (Dijksterhuis et al., 1996).

Is Shared Expertise Most Important for CAT Reviewers?

A multiplicity of creativity raters from diverse backgrounds may be important when using the CAT and DOG because that diversity will follow a sounder sociocultural ontology (Author 3, 2020). Though the CAT claims to take context into account, it ultimately forces consensus, which potentially privileges one perspective of expertise on the judging of artifacts within a domain. The approach to rating the creativity of an artifact is rooted in the life

experience of different evaluators. In the DOG, the goal isn't to make creativity relative (residing, as it were, in the 'eye of the beholder'), but to leverage divergence and disagreement and substantiate the personal and sociocultural nature of creativity in real, everyday contexts.

Looking at the characteristics of the groups of students whose artifacts resided in the ambiguous categories, it is compelling that the most culturally, racially, and ethnically diverse group of students (70% were non-white) was the group whose artifacts demonstrated the highest degree of difference and divergence among raters (APC). Two of the students were multiracial, four students were identified as Hispanic—which itself represents a diverse diaspora of Latin American culture—and one student was Pacific Islander/Native Hawaiian. The raters were all white, so that finding suggests the potential for bias in the ratings in the form of privileging sameness. More specifically, artifacts may have tended to be rated unambiguously creative because they were produced by students who had similar socio-cultural and historical backgrounds. Additionally, the UC group with the highest ratings were 90% female and the raters were 67% female, which further suggests another potential source of “sameness bias” in the ratings.

These issues raise several questions. What are the consequences to our understanding and representations of creativity when researchers rely on panels of rather homogenous raters to evaluate the creative ideas and intentions of students from different sociocultural and historical backgrounds? One problematic consequence suggested from our findings applying the DOG elaboration was raters tended to privilege sameness with respect to features of their own identity and students' socio-cultural and historical identities. This finding highlights the importance of the DOG elaboration for identifying representations of creative expression that would otherwise be eliminated or ignored as statistical noise in the prototypical use of the CAT methodology.

Along similar lines, what are the consequences of using only “arts educators” to make judgments about creative artifacts, as opposed to including teachers representing other subject areas as raters of artistic creations and captions of students? Although domain specificity has long been argued as an important feature of creative expertise in general and assessment of creative artifacts using the CAT in particular (Amabile, 1982; Baer & McKool, 2016), the DOG methodology raises important questions as to broadening the pool and perspectives of raters from different vantage points and experiences. Should adults in any domain be considered the only “experts” of the creative ideas and products of adolescent creators? How does the DOG methodology expand and raise questions of the validity of traditional CAT approaches? More specifically, what are the implications of relying on panels of adult experts, who tend to be homogenous with respect to their socio-cultural and experiential background, to evaluate the products of adolescents who represent diverse socio-cultural and experiential backgrounds? What are potential questions and issues raised for making judgements about creativity when there are imbalances in the socio-cultural and experiential background (e.g., ethnic, cultural, linguistic, or generational) of raters and the creators? This is not to discount the role of expertise and experience, but rather to highlight how the DOG elaboration raises questions about the role *difference* plays in expanding perspectives on how creativity is represented, evaluated and recognized in and across multiple audiences. Perhaps future use of the CAT for the creative products of diverse children and adolescents should include the expertise of young people in the team of raters.

Implications and Concluding Thoughts

Results reinforce the importance of integrating a mixed method approach into the study of creativity in any context. The DOG is a mixed method approach, because it involves the

quantitative ordering process based on the standard deviation of ratings for each artifact, and a qualitative analysis of ambiguously (un)creative production. Measurement approaches to study creativity have started to evolve to become more ecologically valid for creativity in the real world, and the DOG approach represents an important aspect of this evolution. Indeed, creativity researchers need to continue to develop, refine, and use new methods to study the phenomenon of creativity to understand the value and role that disagreement and divergence plays in creative expression and assessments of creative expression. More work will be needed to understand the validity the DOG dimensions of divergent, open-ended, and generative for different contexts and purposes, potentially adding new dimensions to help uncover additional sources of ambiguity.

Our aim in this paper was to introduce and highlight the potential use of the DOG elaboration. We therefore endeavored to raise questions about prototypical techniques such as the CAT and provide examples to illustrate the need for elaborations such as the DOG. We therefore did not provide detailed analysis of ambiguously uncreative and unambiguously uncreative artifacts or artifacts that might fall in the middle of the distributions as average creative and ambiguous (the grey area in Figure 1). Rather, we focused on building the case of the DOG elaboration, how it can be used, and the various categories that the DOG elaboration provides researchers for secondary and tertiary analysis. We encourage researchers to integrate, build on, and refine the DOG elaboration and to incorporate and further develop the analytic guide in Table 3 that emerged from this particular context. Though some of the indicators in Table 3 may be relevant with other types of artifacts in other domains, it is likely that the qualitative aspect of the DOG will need to be both deductive and inductive to identify a comprehensive set of sources of ambiguity and complexity.

The DOG approach also illustrates how important it is for raters to justify their ratings or document the sources of ambiguity that arise for them to be able to remain self-reflexive in the evaluative process of the CAT. In this way, the act of evaluating creativity becomes itself a creative sensemaking process rather than the instantaneous assignment of scores. Debate and deliberation when it comes to judging creativity are important phenomena that we know little about when using the CAT alone.

Last but not least, our proposed methodological elaboration reveals potential biases toward sameness and convergence in assessing creativity in educational contexts. If we are indeed to live up to the standards of an education that cultivates creative experiences—rooted in open-endedness, non-linearity, pluri-perspectivism, and orientation towards the future (Author 3 & 2, 2021)—then we need to embed diversity and divergence in more than our lessons; they need to be reflected in how we discover what is creative, and especially what is potentially creative, and help it grow.

References

- Author 1 (2018a). Withheld for blind review.
- Author 1 (2018b). Withheld for blind review.
- Author 1 (2021a). Withheld for blind review.
- Author 1 (2021b). Withheld for blind review.
- Author 2 (2006). Withheld for blind review.
- Author 2 (2016). Withheld for blind review.
- Author 2 (in press). Withheld for blind review.
- Author 2 (in press). Withheld for blind review.
- Author 2 (2021). Withheld for blind review.
- Author 3 (2010). Withheld for blind review.
- Author 3 (2012). Withheld for blind review.
- Author 3 (2020). Withheld for blind review.
- Author 3 (2021). Withheld for blind review.

- Acar, S., Burnett, C., & Cabra, J. F. (2017). Ingredients of creativity: Originality and more. *Creativity Research Journal*, 29(2), 133-144.
- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997-1013.
- Baer, J., & McKool, S. S. (2016). Assessing creativity using the consensual assessment technique. *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education*, (April), 65–77. <https://doi.org/10.4018/978-1-60566-667-9.ch004>
- Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (Re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 233–240. <https://doi.org/10.1037/aca0000233>
- Bruner, J. S. (1962). The conditions of creativity. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking: A symposium held at the University of Colorado* (pp. 1–30). Atherton Press.
- Bruner, J. S. (1979). *On knowing: Essays for the left hand*. Cambridge, MA: Harvard University Press.
- Clapp, E. (2016). *Participatory creativity: Introducing access and equity to the creative classroom*. New York: Routledge.
- Corazza, G. E. (2016). Potential originality and effectiveness: The dynamic definition of creativity. *Creativity Research Journal*, 28(3), 258-267.
- Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.
- Daker, R. J., Cortes, R. A., Lyons, I. M., & Green, A. E. (2019). Creativity anxiety: Evidence for anxiety that is specific to creative thinking, from STEM to the arts. *Journal of Experimental*

Psychology: General.

Dijksterhuis, A. P., Van Knippenberg, A. D., Kruglanski, A. W., & Schaper, C. (1996).

Motivated social cognition: Need for closure effects on memory and judgment. *Journal of Experimental Social Psychology*, 32(3), 254-270.

Furnham, A. & Marks, J. (2013). Tolerance of ambiguity: A review of the recent literature. *Psychology*, 4(9), 717–728.

Ganesh, T. (2011). Children-produced drawings: An interpretive and analytical tool for researchers. In E. Margolis and L. Pauwels (Eds.). (2011). *The SAGE handbook of visual research methods* (pp. 214–240). Sage

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, 7(2), 193-208.

Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, 49(3), 234-244.

Hennessey, B. A., Kim, G., Guomin, Z., & Weiwei, S. (2008). A multi-cultural application of the consensual assessment technique. *The International Journal of Creativity & Problem Solving*, 18(2), 87-100.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14-26.

Karwowski, M., & Barbot, B. (2016). Creative self-beliefs: Their nature, development, and correlates. In J. C. Kaufman and J. Baer (Eds.), *Creativity and Reason in Cognitive Development* (pp. 302–326). Cambridge University Press.

- Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior*, 43(4), 223-233.
- Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal*, 20(2), 171–178. <https://doi.org/10.1080/10400410802059929>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Long, H., Kerr, B. A., Emler, T. E., & Birdnow, M. (2022). A Critical Review of Assessments of Creativity in Education. *Review of Research in Education*, 46(1), 288–323. <https://doi.org/10.3102/0091732X221084326>
- Makel, M. C., & Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1), 27–29. <https://doi.org/10.1037/a0035811>
- Ness, I. J., & Glăveanu, V. (2019). Polyphonic orchestration: The dialogical nature of creativity. In R. A. Beghetto & G. E. Corazza (Eds.), *Dynamic Perspectives on Creativity* (pp. 189–206). Cham: Springer
- Ness, I., & Søreide, G. E. (2014). The room of opportunity: Understanding phases of creative knowledge processes in innovation. *Journal of Workplace Learning*, 26(8), 545–560.
- Runco, M. A. (1991). *Divergent thinking*. Norwood, NJ: Ablex Publishing Corporation.
- Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, 24, 66–75.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. Retrieved from

<http://www.ncbi.nlm.nih.gov/pubmed/18839484>

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, 27(1), 1–40.

von Hippel E. (1988). *The sources of innovation*. New York, NY: Oxford University Press.

Zenasni, F., Besançon, M., & Lubart, T. (2008). Creativity and tolerance of ambiguity: An empirical study. *Journal of Creative Behavior*, 42(1), 61–72.