Design Creativity Assessment and Capturing

The Third International Conference on Design Creativity (3rd ICDC)
ISBN 978-1-904670-60-5
Bangalore, India, 12th-14th January 2015

A CAT WITH CAVEATS: IS THE CONSENSUAL ASSESSMENT TECHNIQUE A RELIABLE MEASURE OF GRAPHIC DESIGN CREATIVITY?

K. K. Jeffries[1]

[1]School of Art, Design & Performance, University of Central Lancashire, Preston, UK.

Abstract: The CAT is considered one of the gold standards for creativity assessment, and graphic design, arguably, the most ubiquitous domain within the "creative industries". For the first time, this study tests two tasks to measure graphic design creativity, and by extension, the reliability of the CAT as a measure of graphic design creativity. Initial research suggested low inter-rater reliability may be unduly influenced by a judge's preference for technical execution. Sixteen professional graphic designers were randomly assigned instructions to discount technical execution from creativity ratings, or instruction that gave no stipulation, for 60 artworks. Inter-rater reliability were acceptable for each task and experimental condition, but were higher for judges that received instructions to discount technical execution. These and other results are discussed, and the argument presented that, for future CAT studies in this domain, specific instructions to discount technical execution offers a more reliable measure of graphic design creativity.

Keywords: *Consensual Assessment Technique, Graphic Design, Design Creativity Assessment*

1. Introduction

Within the "Creative Industries" graphic design links across many sectors: be it the need for marketing material; a new logo for an organization; the presentation of scientific information, or the development of a new product. Graphic design will play a part: sometimes in the background, at other times centre stage.

Creativity, in this context, is an asset, and gives clients, design agencies, and individual graphic designers an edge in a competitive market. Whilst not all design opportunities give scope for creativity, many do, and finding novel ways to visually communicate with an audience is often an implicit expectation for competency. Few would argue assessing graphic design creativity is anything other than a highly subjective process; however, such subjectivity need not be problematic if assessors concur in their subjective judgments. This is exactly what the Consensual Assessment Technique (CAT) maintains to achieve; a research method that can reliably assess creativity, through the consensual assessments of domain experts.

Despite extensive use within creativity research, the use of the CAT as a measure of creativity within design research is relatively small. Over a thirty year period, for example, only 11 papers are related to design journals (Jeffries, 2012), and on further review, only two papers operationalized the CAT in their research. Building upon this thirty year CAT database, follow up searches, identified twenty four papers that made reference to both graphic design and the CAT. On inspection, the vast majority of studies, however, were not related to graphic design directly, with two exceptions (Silvia et al., 2008; Dineen & Niu, 2008).

Silvia et al (2008) undertook a study to validate a new method of scoring divergent thinking tasks. Part of the study had participants who majored in arts subjects (accounting for 9% of participants in the study), of which some participants majored in graphic design. Whilst the new technique built the case for the validity of subjective rating by citing CAT studies, the method used was not the CAT protocol. Indeed, they acknowledged the importance of expert domain judges for "studies of real creative product", but argued that this need not be the case for divergent thinking tasks assessment.

In contrast, Dineen and Niu's work (2008) utilised the CAT method, and involved participants in their 2nd year of graphic design at a Chinese art and design higher education institution. The study explored the respective merits of UK creative pedagogy relative to traditional Chinese pedagogy. It is arguable, however, how closely aligned to graphic design the final outputs were relative to illustration. In no way do distinctions that can be made between graphic design and illustration design undermine the value of Dineen and Niu's research, but for the purposes of this review it highlights that at the time of this study, there was at best only one published paper that had applied the CAT to graphic designers.

Therefore, as yet, there is little precedent established for a task to measure graphic design creativity using the CAT. Particularly in the light of debates surrounding the domain specificity/generality of creativity, and the role task selection plays in creativity assessment (Byrne, 2011), this raises both practical and theoretical implications for creativity research within and beyond graphic design. Thus, for the first time, this study aims to tests two tasks to measure graphic design creativity, and by extension, the reliability of the CAT as a measure of graphic design creativity.

2. Pilot studies

Prior to the main study in this paper, two tasks were piloted, each of which had current usage in design education, and gave participants the opportunity for graphic design creativity. One used text only, and required the graphic designer to choose a word, and then visually communicate that word through the use of type. For example: the word "subtract" could become "Subtrct" or "-tract"; "stop" could become "s t o p!". The other task was to design a graphic which could be transferred to a plain white T-shirt. The graphical image was to be based on the theme of "hands". For brevity, only a selection of pilot findings will be discussed, particularly those that informed the current research design.

A key finding was that inter-rater reliability was below acceptable levels, or marginal for both tasks: 0.56 for the T-shirt task, and even 0.69 for the type task was not ideal. Was something happening in graphic design that warranted caveats for research design, or was this domain challenging the assumptions upon which the CAT was based (i.e. domain experts can independently agree on creativity to an acceptable level of inter-rater agreement)? The outcome from these pilot studies was that a number of caveats needed to be considered in order to optimise the CAT method as a measure of graphic design creativity: specifically, the influence of technical execution on ratings of graphic design creativity; the background of judges; the range of artwork available, and the suitability of a task for research purposes.

2.1 Technical execution

The outcomes of a graphic design brief can vary in its level of technical refinement. At one end of the spectrum are "conceptual" outcomes (these may lack polish and technical detail, but the seed of the idea can be perceived); at the other end of the spectrum is "finished" artwork (ready to go to print or publication). Is it fair to judge a "conceptual" outcome with the expectations held for a "finished" artwork? Likely not, yet, the pilot studies highlighted that judges' preference for technical execution differed, and appeared to be heightened or subdued relative to their interpretation of the design brief and creative outcomes they were judging. Perhaps if graphic design experts are not specifically guided to discount technical execution then ambiguities impact on the level of consensus: such differences in technical preference may explain low inter-rater agreement in previous studies. Within the CAT literature, some researchers have created instruction that directly address this issue of discounting technical execution; other researchers (the majority of CAT studies) have not done so with little adverse impact. In this study, two sets of instructions will be given to judges to test whether discounting technical execution, or not, has an impact on levels of consensus regarding graphic design creativity.

2.2 Suitable judges

In the pilot studies novice/intermediate judges, relative to experts, achieved higher levels of consensus for the type task but the rankings, and rationale for ratings, where not the same. The debate over the use of novice or intermediate judges is a contentious one, and likely to remain as such. For future CAT study in graphic design the implication was to use domain experts only. A further point is that domain experts can be based within professional practice (i.e. full-time graphic designer), or educational practice (i.e. full-time graphic design lecturers teaching on undergraduate and postgraduate graphic design courses). Previous research in other design domains (Jeffries, 2009) suggest that the values shared on creativity between academics and practitioners is not as polarised as populist views can imply. However, whether this is the case for graphic design is unclear, and thus a cautious approach to CAT studies, in this domain, would be to gather experts from either academia or professional practice, but not to use both within the same group (that is, until research findings can show otherwise). In the present study, only full-time professional graphic designer were used as judges, and each judge was required to have over two years of professional experience within graphic design.

2.3 Range of artwork

It may also be the case that the artwork sampled for the pilot studies lacked diversity and this had an impact on inter-rater reliability. Whilst the CAT is a method that highlights ratings be relative to other works within a sample, it is feasible that artworks too similar in quality pose a more challenging task for judges than those that show more diversity. Such a possibility is interesting. Several CAT studies highlight the real world basis of the technique, and with acceptable inter-rater reliability there is little reason to question further. However, most CAT studies are research studies, and whilst parallels can be drawn between CAT methods and those used by judges of, for example, professional competitions, awards, or degree show assessment, the independence of judges does not happen throughout the rating process as it does in the CAT. Competition judges do tend to confer with each other towards the end of the process. This need not discredit CAT methodology, but it does suggest that the real world assessment of highly creative artworks may require more debate amongst judges than the CAT allows. If this is the case for highly creative artworks, then why not for other skewed samples, be they predominantly low or medium in quality. It is beyond the scope of this research to explore this in greater detail, but the argument here, given the pilot findings, was that some form of pre-test for diversity of artwork was prudent.

2.4 Task selection

As a final consideration, the T-shirt task may have been too complicated for the purposes of this study. Firstly, as a task it was designed for an educational purpose with a broad scope, and one that took place over an extended period of time (relative to the more experimental tasks used by many

CAT studies). Secondly, the artwork ranged from hand-drawn sketches to "Mac'd" up. Moreover, and thirdly, some artworks incorporated text within the T-shirt graphic, whilst other did not. Each of these issues may have contributed to the low level of consensus achieved in the pilot study. Given this, the implication was for artworks to contain image only or text only, and that artwork would be either all hand-drawn or Mac'd up (but not contain a combination of both). Additionally, the task should be able to be completed in a time scale comparable to the type task.

Based upon such caveats, the main study was specifically designed to ask: can professional graphic designers achieve inter-rater reliability at or above 0.7 for a image based graphic design task and a text based graphic design task? If they can, do CAT instructions to discount technical execution increase inter-rater reliability when compared to instructions that make no such stipulation?

3. Method

As is often the custom for CAT studies, two broad groups of participants were required: those who generate the creative outputs (participants), and those who assess the creative outputs (raters, or judges). For many CAT studies, participants are recruited and undertake the creation of a piece of work under experimental conditions -this is because the purpose of such studies is to test the influence of teaching or environmental factors that may impact upon creativity. However, in this study, it is the judges, rather than the participants that are the focus. In this respect, this study follows the research direction set by Baer, Kaufman & Gentile (2004) for the use of work created under non-experimental conditions. In this work they showed that judges' inter-rater reliability remained acceptable even when the creative outputs were not generated under experimental conditions.

For this study, graphic design artwork was the creative output, and created as a natural result of engagement with a university degree. Specifically, the study gained consent from participants to use type only and image only artwork created during two assignments for a BA (Hons) Graphic Design course. These two tasks provided 30 type and 30 image examples to be independently assessed by 16 professional graphic designers using the CAT. The dependent variable was instructions given to judges, and each judge was randomly assigned to receive different instructions for each task. Data was analysed for inter-rater reliability, and appropriate statistical analysis was used to compare the influence of different instructions on judges ratings. As this project involved two universities, authorisation was sought, and obtained, from ethics committees at each.

3.1 Tasks

The type task was the same as the pilot study. A new image task was used that required participants to select two images, and when seen side by side made some sort of creative juxtaposition, or interesting visual communication about the images chosen. Like the type task, this has a pedigree within graphic design education, and the juxtaposition of images is also a technique frequently used by professional graphic designers.

3.2 Instruction to judges

For each task, whether type or image, two sets of instructions were developed. One set was an adapted version of Kaufman, Baer, Cole & Sexton's 2008 study, and is cited as an exemplar of CAT instructions (Kaufman, Plucker & Baer, 2008).

"Please look through these artworks, and rate them for creativity. There is no need to explain or defend your ratings in any way; we ask only that you use your own sense of which is more or less creative (relative to the other artworks provided).
Please look through these artworks three times, and rate them for creativity.
The first time familiarize yourself with all the artworks provided.

The second time, group the artworks into Low, Medium, or High ratings.
The third time, assign a numerical rating between 1 and 6 (1's being the least creative and 6's being the most creative).
There should be a roughly even number of artworks at each of the six levels. It is very important that you use the full 1-6 scale."

The other set of instructions were exactly the same as Kaufman et al's, but inserted was a sentence adapted from Baer's 1993 CAT instruction for judges (a study where judges where specifically requested not to consider other factors that may impact on storytelling, for example, aesthetic appeal, or grammar, as part of their criteria for rating creativity). Baer's (1993) study is also cited as an exemplar of CAT instruction (Kaufman, Plucker & Baer, 2008).

"We realize that creativity probably overlaps other criteria one might consider (for example: aesthetic appeal, or technical execution) but we ask you to rate the artworks solely on the basis of their creativity."

3.3 Selection of Artwork

In order to select 30 artworks for each task, grades created as a result of student artwork for a creative thinking module were used: the criterion was creativity, and this was assessed by academic staff independent of this study. The details of this selection process have been reduced for brevity, but the purpose was threefold: firstly, to identify a diverse range of artworks across all CAT levels 1 to 6; secondly, to have five artworks represented at each CAT level; thirdly, to have the same participant represented in both tasks: type, and image. This would allow CAT scores for type and image to be aggregated to get a total score for both tasks and individuals. As above, the CAT ratings for this study would require judges to rate artwork as low, medium, or high, and then rate these from 1 to 6. To determine which artworks, and participants would be selected, each academic grade was stratified to a CAT rating as follows: marks between 44 and below, and 44-50 where rated as low (CAT level 1, and 2); 50-54, and 55-59 as medium (CAT level 3, and 4); 60-69, and 70 and above as high (CAT level 5, and 6). When each of the artwork options were placed alongside each other, this highlighted some CAT levels had fewer options than others. For example, only five artworks were available at CAT level 1 for the type task, and only five artworks were available at level 6 for the image task. The inclusion of a participant in the type task at CAT level 1, for example, determined their representation within the image task, and vice versa. Moreover, when a participant was chosen to represent a specific CAT level (regardless of which task), this influenced the options available for other CAT levels. In this respect, the choices for selection became to an extent self-identifying, with limited options depending on whether a CAT level had more than five artworks available. The process was iterative, and became progressively more challenging with each inclusion. However, the stratification was achieved, and 30 participants were identified whose text and image artworks represented each of the six CAT levels, with five artworks at each level. For reasons of research ethics and participant confidentiality, examples of the artwork used in this study have been withheld from publication.

3.4 Judges

Previous research guidelines suggest that for CAT reliability, between 5 and 10 judges is an acceptable number for a given task. For this study, 16 full-time professional designers took part (sampling detail are available in the results section). Judges were randomly assigned to two groups, in which task and instructions where rotated to counter balance whether they receive the image task first or second, with whether they received Kaufman or Baer instructions, or a combination. This was to minimize order effects, particularly practice effects and fatigue effects.

3.5 Procedures

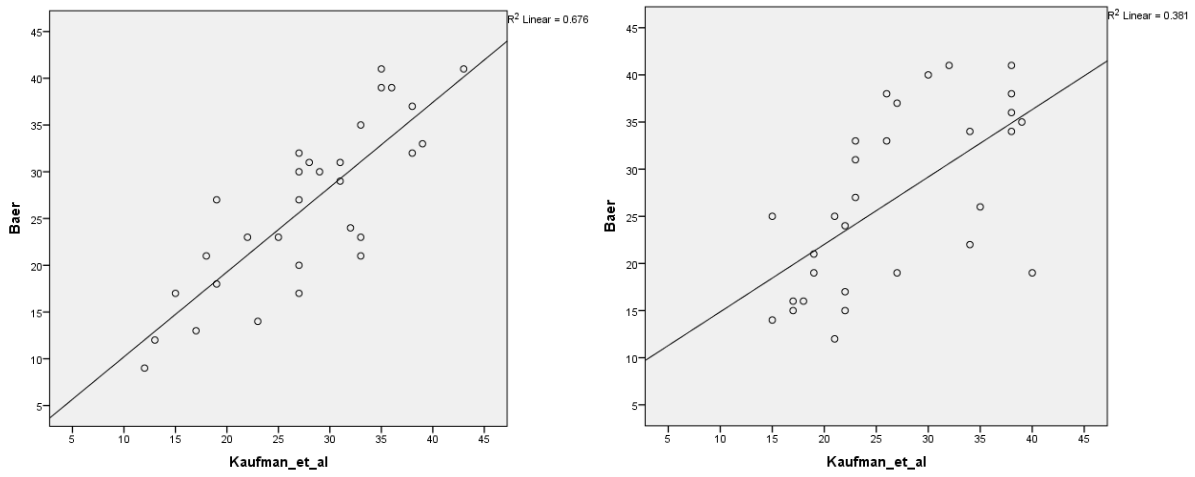
The procedures for rating artwork were the same for each judge, and task. Initially, judges answered three questions: years of experience in graphic design, whether they would describe themselves as a

graphic designer and their age. After this, each judge was given the instructions for their first task. They had as much time as they required to read the instruction. Next, each judge accessed a laptop with a PDF slide presentation of the 30 artwork for the first task. The order of artwork was randomised, and they were free to control how long they viewed artworks, and could return to each artwork for further inspection. Each judge familiarised themselves with all the artworks, and when satisfied informed the researcher they were ready to continue. Judges were given an A3 laminated rating sheet, and a set of laminated cards. These cards were miniature copies of the artwork they had just viewed. Cards were placed in a stack, by the researcher, onto the rating sheet and judges proceeded to rate the artwork, and had as much time as they required. Task two followed the same procedures. All judges were debriefed on the purpose of the experiment, and had the opportunity to ask any questions about the study.

4. Results

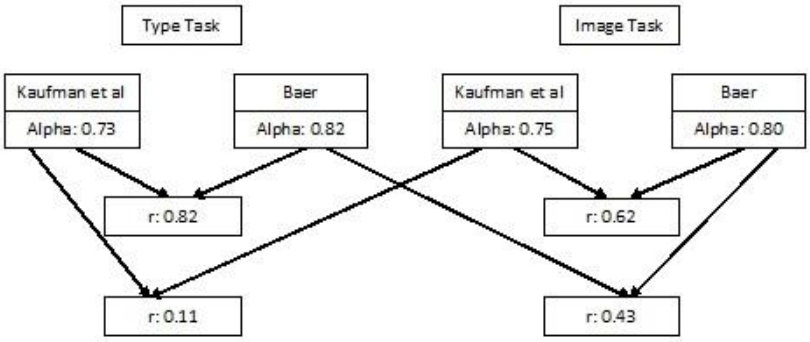
Within a year group of 66 students, 48 students gave their consent to take part in the project. The median age was 19 years (SD 1.46); 18 female and 30 male students took part. The mean age for judges was just over 41years of age (SD. 9.80), and ranged from 30 to 63 years of age; six judges were female. All judges identified themselves as graphic designers, and their professional experience within graphic design ranged from 7 to 35 years: the mean being just over 17 years.

For the type task, the 8 judges who received the adapted Kaufman et al CAT instructions had an alpha of 0.73; for the 8 judges that received the adapted Baer's instruction to discount technical execution the alpha was 0.82. Pearson's r was 0.82, suggestive of a very strong positive correlation between the scores, and was a significant correlation at the 0.01 level (two tailed). The scatter plot on the left shows the regression line, and strength of correlation.



For the image task, the 8 judges who received the adapted Kaufman et al's CAT instructions had an alpha of 0.75; for the 8 judges that received the adapted Baer's instruction to discount technical execution the alpha was 0.80. Pearson's r was 0.62, suggestive of a strong positive correlation between the scores, and was a significant correlation at the 0.01 level (two tailed). The scatter plot to the right shows the regression line, and strength of correlation.

The correlation between the adapted Baer's instructions for both type and image was an r of 0.43, which can be considered a moderate positive correlations, and one that was significant at the 0.5 level (two tailed). For the adapted Kaufman et al instructions for both type and image r was 0.11, suggestive of a negligible positive correlation, and one that was not statistically significant.



5.Discussion

Given the novelty of the CAT for the assessment of graphic design creativity, there was a need to pilot whether judges showed a suitable standard of consensus as achieved in other research. Of the few studies directly related to design, the CAT has shown sufficient levels of consensus within their respective domains. It was expected that using the CAT to assess graphic design creativity should follow a similar pattern of inter-rater reliability; however, the pilot result did not achieve this. A number of reasons, such as the range of artwork, technical preferences of judges, task selection, and the sampling of judges were accounted for in the present study, and appear to have resolved the previous issue of inter-rater reliability. In either task, the inter-rater reliability was acceptable, all were above 0.7, the highest being 0.82. Given this, these particular tasks can be considered reliable measures of graphic design creativity using the CAT. Prior to this study, the choice of tasks to measure graphic design creativity was not obvious, and the pilot findings highlighted that some task do not translate well from design education context into experimental research.

Whilst each task (regardless of the instruction to judges) had acceptable alphas, a marked difference can be seen in the correlation between the type task (r=0.82) and the image task (r=0.62). The reasons for this difference could be numerous, complex and interrelated. Perhaps the type task, and typography, shares a common knowledge base for graphic designers; possibly the type task is less influenced by discounting technical execution than the image task; the inclusion of colour in the image task may add to the complexity of assessing its creativity: colour was absent in the type task. What can be said, and has been said by other researchers (Reiter-Palmon et al, 2009), is that task selection is an important factor in creativity assessment, and our depth of understanding is "essentially missing in the literature" (Lubart & Guigard, 2004, p.48)."

In early CAT research, Amabile (1996) concluded that judges were able to distinguish creativity from other aspects such as aesthetic appeal and technical execution. Does this finding still apply to graphic design creativity? Whilst the differences between the adapted Kaufman et al instructions and the adapted Baer instructions (to discount technical execution) ranged from 0.05 to 0.09, the difference was towards higher levels of inter-rate reliability when judges were asked to discount technical execution from their creativity ratings; this occurred in both the type task and image task. Arguably, these differences are slight, but it may be that acceptable inter-rater reliability is not enough in isolation, and that the other consideration is the correlation between aggregated scores. Only the adapted Baer instructions were statistically significant, and suggestive of a moderate to strong positive relationship. A further finding was that correlations were significant at the 0.01 level in either task, but were stronger for the type task, than the image task. It appears the type task is less influenced by discounting technical execution, and the image task may be more susceptible.

The main point to consider is whether to include a caveat around technical execution in future research. By inclusion, such a caveat, directly addresses assumptions around technical execution and creativity. Indeed, if the CAT is foremost a measure of creativity (however judges interpret this

word), then clarification on technical execution seems a reasonable distinction to bring to their attention. The slight increase in inter-rater agreement for judges that received the technical execution caveat can be interpreted both for and against its inclusion. More revealing is the correlations between type and image scores relative to instructions. It is only scores where a caveat was included that enabled an aggregated graphic design creativity score. As the purpose of these tasks was to evaluate graphic design creativity by isolating two distinct features of graphic design (the creative use of type, and the creative use of image), the expectation, however, was that a degree of positive correlation would be likely in the combination of these tasks. What is interesting to note for this study is that when judges assess exactly the same tasks and exactly the same artwork, only the caveat on technical execution offers the opportunity for an aggregated score, and thus, at least for research purposes, enables distinctions within a group on levels of graphic design creativity.

6. Conclusion

Prior to this study the choice of tasks to measure graphic design creativity was not obvious. In this study a number of research design factors, such as diversity of artwork, technical preference, task selection, and sampling of judges were accounted for in the research design, and appear to have resolved previous issues of inter-rater reliability. However, the arguments presented in this paper suggests instructions to discount technical execution from judges' creativity assessment do appear to influence the reliability of the CAT. The difference was towards higher levels of inter-rate reliability when judges were asked to discount technical execution from their creativity ratings; this occurred in both the type task and image task used. Moreover, only CAT assessments undertaken where the technical caveat was included enable an aggregated graphic design creativity score for both the image and type tasks. Perhaps these implications apply not only to graphic design, but have relevance for all CAT assessments of design creativity? To paraphrase Nickerson, as researchers we have two choices: include a caveat on technical execution that future research will show was not fundamental, or exclude it and find technical execution does influence rating on design creativity. Unless there is some detrimental effect (which does not appear to be the case) then a cautious approach would be for future CAT usage to include a caveat on technical execution when applied to design creativity research.

References

Amabile, T. M. (1996). Creativity in context: Update to the social psychology of creativity. Westview Pr.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the Consensual Assessment Technique to Nonparallel Creative Products. Creativity Research Journal, 16, 113-117.

Baer, J. (1993). Creativity and divergent thinking: A task-specific approach. Lawrence Erlbaum Associates, Inc.

Byrne, C. (2011). Task selection and the consensual assessment technique: Using collage tasks in creativity research . University of Central Lancashire.

Dineen, R. & Niu, W. (2008). The effectiveness of western creative teaching methods in China: An action research project. Psychology of Aesthetics, Creativity, and the Arts, 2, 42.

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). Essentials of creativity assessment. Hoboken, NJ US: John Wiley & Sons Inc.

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, 171-178.

Jeffries, K. K. (2012). Amabile's consensual assessment technique: Why has it not been used more in design creativity research? Proceedings of the 2nd International Conference on Design Creativity (ICDC2012), Vol. 1, 211-220.

Jeffries, K. K. (2009). Skills for Creativity in Games Design (Part 1): Academic Conceptions of Creativity in Games Design. Brighton: Higher Education Academy: Art, Design and Media subject centre.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I. et al. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2, 68.