Design Creativity, Technical Execution and Aesthetic Appeal: A CAT with Caveats (Part 2)

This study explores to what extent technical execution and aesthetic appeal may be related to assessments of graphic design creativity. These new research findings build upon Jeffries' 2015 publication in the International Journal of Design Creativity and Innovation, and further underpin the caveats identified in relation to the Consensual Assessment Technique (CAT). Eight professional graphic designers rated thirty-two artworks for a creative typographical task. Individual artworks were created by novices who had no experience of graphic design, through to professional graphic designers with 35 years of full-time experience. Written instructions to judges emphasised artwork be rated on creativity only (without considering technical execution or aesthetic appeal), and this "creativity only" feature was verbally re-emphasised to judges by the researcher. Inter-rater agreement for creativity was a Cronbach's alpha of 0.92; considerably higher than in previous studies, with implications that may relate to the use of the CAT as a measure of design creativity more broadly, and beyond graphic design.

Keywords: Consensual Assessment Technique (CAT), graphic design, design creativity assessment

1. Introduction

"...A product or response is creative to the extent that appropriate observers independently agree it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated." (Amabile, 1982, p. 1001).

Amabile's description of creativity measurement through social consensus by domain experts has formed the theoretical framework for thousands of research projects. In turn, the Consensual Assessment Technique (CAT) has been shown by many researchers to offer a reliable and valid operational definition of creativity (Baer & McKool, 2009). In

its practical application, CAT guidance states that when creativity is assessed in a new domain (i.e. one that has not been studied with a particular task), researchers should ask judges to rate additional constructs, specifically, technical execution and aesthetic appeal (Amabile, 1982; Hennessey, 1994), and check to see if creativity ratings are distinct from these criteria.

For example, artworks in graphic design can vary in their level of technical refinement. At one end of the spectrum are conceptual sketches, where the seed of an idea can be perceived: even if the sketches lack refinement in, for example, font selection, layout or composition; the creativity of the idea can still be evaluated. At the other end of a technical spectrum are finished artworks; artworks that are ready to go to print or publication, where every aspect of visual communication has been crafted and refined to perfection by the designer.

In regard to CAT protocol, once the relationship between creativity, technical execution and aesthetic appeal has been shown to be distinct for a given task, researchers need only ask for ratings of creativity for subsequent studies and can assume that technical execution and aesthetic appeal are no longer a conflating issue for validity. Indeed, in Amabile's 1982 work the extent to which creativity may be isolated from such factors was a formative part of her paper, and she concluded that "...although judges were not provided with a definition of creativity...they consistently and reliably identified a quality in both types of product that was distinct from technical execution. Moreover, for artworks, it was distinct from aesthetic appeal as well" (p.1010).

Hennessey's discussion of the CAT (1994), however, acknowledged that for "some domains" (p.195) the distinction between technical execution and aesthetic appeal may be less clear and that creativity is likely to correlate with these aspects. Indeed, in

several design related studies evidence can be found of this. For example, Christiaans' (2002) applied the CAT to Industrial Design Engineering students, and found strong correlations between creativity and aesthetic criteria, but not for creativity and technical quality. Wojtczuk and Bonnardel's (2012) use of the CAT found a positive correlation between creativity and four other criteria (aesthetics, originality, brief appropriateness and audience appropriateness) for a poster design task.

This study explores to what extent technical execution and aesthetic appeal may be related to assessments of graphic design creativity. The results are relevant to graphic design but also extend to other areas of design creativity, and consideration of CAT protocol more broadly.

2. Background

Within the Creative Industries graphic design links across many sectors: be it the need for marketing material; a new logo for an organisation; 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.

Established within traditional print media alongside the development of the Digital Creative Industries, the roots of graphic design, arguably, do not first begin in the 20th century. Visual communication: turning words into icons, or using images *to say a thousand words* is ancient. Like many established professions its boundaries blur and change, and the line between graphic design, illustration, advertising or copywriting is debatable and in flux (we could add other domains to this list: typography, printing, web design, photography, animation, packaging, let alone new technologies searching to establish an identity).

Yet, within this flux, graphic design exists and is global. Throughout the world, universities, colleges and private companies run graphic design courses; their students' gain qualifications that enable them to work for design agencies or set up their own business, and they in turn work for clients who require graphic design to appeal to their audience and customers. In this context, creativity is an asset and gives clients, design agencies, and individual graphic designers an edge in a competitive market.

Since the first systematic review of the CAT in design publications in 2012 (Jeffries), CAT has continued to follow an upward trend in growth within scholarly journals. Indeed, more CAT studies and citations occurred in the five-year period between 2010-2015 than in the previous 28 years of CAT research. The impact is that a greater number of researchers, both within and outside of design, apply CAT as their operational definition of creativity.

However, despite Amabile's original guidance, and subsequent guidance by Kaufman, Plucker, and Baer (2008) who dedicated a complete chapter of their book *Essentials of Creativity Assessment* to the CAT, as did Hennessey, Amabile, and Mueller (2011) for the *Encyclopaedia of Creativity*, the way different researchers implement the CAT in practice shows substantial inconsistencies (Cseh, 2014). To what extent differing interpretations of the CAT protocol influence the reliability and validity of the method remains a topic for international debate, and this paper continues to explore such considerations in relation to design creativity, and graphic design specifically.

Prior to 2015, data from the use of CAT in graphic design were not widely published. Once these data became available, it was clear that a graphic design CAT appeared

more sensitive to methodological protocols than previous studies in other domains had reported. Access to these results served as an encouragement to understand further what may be influencing these low levels of consensus and to revisit past assumptions upon which the CAT was established. This paper explores three interrelated assumptions around the effect of "creativity only" instructions, the level of correlation between creativity, technical execution and aesthetic appeal, and the diversity of artwork presented to judges.

CAT ratings are based on the normative assessment of creativity within a defined group (i.e. the specific participants in a study). As a result, Jeffries (2015) argued that diversity of artwork may play a significant factor in levels of consensus amongst judges, and researchers should not assume that there is enough diversity in a sample of artwork (particularly if the consensus is low).

How such diversity of artwork is achieved can occur in several ways. For example, Jeffries' 2015 study used pre-selection of artwork based on academic grades; other researchers, like Christiaans and Venselaar (2005), randomly sampled from a population of 240 students to select 55 and then used higher and lower creativity scores to select 18 cases for qualitative analysis.

As an alternative to randomization or pre-selection, in the present study, participants were stratified according to levels of expertise (novice, intermediate and expert) within graphic design. The logic for stratification was based on the assumption that experts are likely to produce more creative graphic designs than intermediates, and intermediates more than novices. While in specific instances that assumption may not hold, it is the

group as a whole that forms the relative basis as a measure of creativity, and stratified sampling offers a means to achieve this diversity of artwork.

2.1 Research questions and hypothesis

The study used a creative typographic task that was developed in Jeffries (2015). This Type Task was based on the seminal graphic design book Watching Words Move (Chermayeff & Geismar, 2006) and, as a design exercise, has longstanding and current usage in design education. The Type Task uses text only, and requires the participant to choose a word, and then visually communicate that word through the use of type. For instance, in figure 1 the word "Saw" is written by extending the last letter "W" to suggest the teeth of a saw, or the word "Imagine" has the middle letter deleted, leaving the viewer to imagine what this deletion may be.

Figure 1. Two examples of the artwork related to the Type Task (neither example was part of this study)

Building on the 2015 study, the current study was guided by two research questions:

- Can the Type Task remain a reliable measure of graphic design creativity without pre-selection of artwork?
- To what extent are technical execution and aesthetic appeal correlated with graphic design creativity?

The expectation (given the finding of Jeffries, 2015) was that Cronbach's alpha for graphic design creativity would remain above 0.7, and, based on previous CAT studies focused on design creativity (such as: Christiaans, 2002; Wojtczuk & Bonnardel, 2012; Valgeirsdottir, Onarheim & Gabrielsen, 2015), that there would be a moderate to strong significant correlations between judges rating of technical execution, aesthetic appeal, and creativity, in most, if not all, instances.

3. Method

CAT studies require two broad groups of participants: those who would generate the creative outputs (participants), and those who would assess the creative outputs (judges). Creative output (in this case, graphic design artwork) was judged by eight professional graphic designers, and a range of statistical analysis methods were used as required, Cronbach's Alpha, Pearson's Correlations Coefficient and Factor Analysis.

3.1 Participants

While participants were stratified between novices, intermediates and experts, to offer a range of graphic design creativity, it was important they were considered as one group, as a whole, for data analysis. For some researchers, thirty participants are considered to be a threshold for the use of parametric statistics (Gall, Borg & Gall, 1996); for others, this is a topic of debate and would prefer to set a baseline of 50 participants. For this reason, a power analysis was performed and highlighted that with a sample size of 32 participants, a one-tailed correlational analysis would have 80% power to detect a moderate correlation if it was present in the data.

Before undertaking the creativity tasks, each participant's level of experience in graphic design was checked through a short questionnaire after they had received a consent

form and information sheet. The study was granted ethical approval on the basis that levels of experience would not be correlated with creativity scores. Each level of expertise was identified by gaining information as follows:

1: I work as a full-time graphic designer, and have done so for more than five years
2: I have worked as a designer, artist or craftsperson in the last ten years.
3: I am currently studying a full-time graphic design degree, and have done so for more than two years
4: I have studied design, art or craft either formally or as a hobby in the last ten years.
5: I have not studied design, art or craft either formally, or as a hobby, since leaving secondary school.

After completion of the study, participants had the chance to ask questions and were given a £10 Amazon voucher as a gesture of thanks for their support. In total 118, professional graphic designers, student graphic designers and members of the general public were contacted. 32 took part and completed the Type Task; 18 females and 14 males, with a mean age of 32.69 years (SD: 11.27), that ranged from 19 to 58 years of age.

Novices: ten females and one male. The mean age was 35.27 (SD: 7.95 years) and ranged from 23 to 49. Seven participants had not studied design, art or craft either formally, or as a hobby, since leaving secondary school. Four participants said they had studied design, art or craft either formally or as a hobby in the last ten years. Example comments for this rating were: "I do a lot of crafts as a hobby. Crochet, jewellery", or "I

have done a little silversmithing and bookbinding as a hobby - more technical than design. No formal study, or art or graphics at all."

Intermediates: six females, and four males. The mean age was 20.70 years (SD: 1.34) and ranged from 19 to 24. All described themselves as currently studying a full-time graphic design degree, and have done so for more than two years.

Experts: two females, and nine males. Experience in graphic design ranged from 4 to 35 years, with a mean of 19 years (SD: 9.77 years). The mean age was 41 (SD: 10.08 years) and ranged from 26 to 58 years. The majority of experts described themselves as a graphic designer, designer or creative; other titles were an art director, creative director, or brand engagement specialist.

The work assessed was created under non-experimental conditions, as it has been shown that CAT inter-rater reliability could remain acceptable even when creative outputs are not generated under experimental conditions (Baer, Kaufman & Gentile, 2004; Christiaans & Venselaar, 2005). The Type Task was completed by participants in their own time, and unsupervised by the researcher.

3.2 The Type Task

The Type Task (see figure 1.) was the same graphic design creativity task used in Jeffries, 2015, with some minor adaptions. Of three groups of participants, novices were the most challenging group in respect to undertaking the Type Task. Experts and intermediates were likely familiar with the Type Task through their professional and educational experience, and would certainly be versed in the computer software that aided undertaking the Type Task. This was unlikely to be the case for novices.

The probable mismatch in technical software skills between novices and intermediate/experts was considered. For example, if only basic computer software were used (Microsoft Word and PowerPoint are software packages available to the vast majority of working participants), while this may match the technical skill level of the novices, it could undermine the creative freedom of the intermediate/experts.

Perhaps of more importance is that the limitations of such software may restrict the creativity for all participants, as the challenge becomes not about one's creative vision, but how to implement this within the constraints of the software. In contrast, to specify that designs be created on specialist software (such as Adobe Illustrator or InDesign) clearly disadvantages novices: they would likely be unfamiliar with such software (which is daunting for most at first) and significantly more of their time would be taken with understanding how to use the software rather than the creativity of their design. What was required was to reduce this mismatch in technical software skills; to the extent that not knowing how to do something in a computer package did not impinge on the creativity of the concept, and its resulting manifestation to the satisfaction of the participant. As two-thirds of the participants would be intermediate or experts, the decision was taken to use Adobe Illustrator, and not to restrict the software they would have available for the Type Task to basic software.

To make sure that novices were not disadvantaged, the solution was to give novices access to software technical support that enabled them to realise their ideas. In this respect, novices may not know how to produce their ideas in Adobe Illustrator but would be able to dictate their ideas to a technician who could. Exactly how this support

was provided, for how long, and how to guard against software support slipping into design suggestions, and thus not reflecting the creativity of the participant, was considered. For example, decisions on layout, type, and choice of work would belong to the novice, with the technical role purely to transfer ideas into Adobe Illustrator. Such conditions were presented to participants through one A4 sheet on technical support; relevant details are highlighted as follows:

As this Type Task is designed to offer a creative challenge across the full range of graphic design expertise, it is important that the tools available to create artwork provide the greatest freedom of expression…The technical support is to enable each participant, regardless of expertise with Adobe Illustrator, to create an artwork that reflects his or her creative vision. It is vital, however, that this technical support only deals with the practical aspects of transferring a participants design to Adobe Illustrator, rather than offering advice or suggestions on the quality of the design…To safeguard against such influence, where possible, technical support will be offered remotely via email. For example, a participant may send a hand drawn illustrations of their design via email to kjeffries@uclan.ac.uk. This hand drawn illustration will then be transferred to Adobe Illustrator, and this digital illustration will be sent back to the participant. Any amendments required by a participant can be suggested, and the cycle will continue until the participant is satisfied, or until the time limit on the task has expired…The time required creating this artwork is expected to take no more than one hour. However, thinking time and playing with ideas may increase for some participants…This task is not timed in a conventional sense but for practical purposes, a limit of one week from receiving the task has been set…Unless otherwise stated by a participant, the default font will be Gill Sans, and the size of the type will be selected to fit an A4 sheet (allowing for a 20 mm border).

Before beginning the recruitment of participants for this study, the Type Task was piloted with three novices. Importantly, the application of technical support did not appear to influence the design a participant had visualised. Through photographs of sketches, and written instructions novice participants were more than capable of expressing their creative vision, and ask for amendments until the design was to their satisfaction.

3.3 Judges and Rating Procedures.

Eight judges independently rated the artwork using the CAT procedures outlined in Jeffries (2015) and discussed further in this section. From these eight judges, four were females, and four were males. The age range was between 31 and 55; the mean was 41.63 years (SD of 8.88 years). Years of experience in professional graphic design ranged from 9 to 34 years; the mean was 19.13 years, (SD of 8.43 years).

After signing the consent form, each judge rated the 32 artworks for creativity only (i.e. discounting technical execution or aesthetic appeal from their creativity rating). Specifically, the creativity only instructions were worded as follows:

Instructions for Judges: How to rate these artworks

There is only one criterion in rating these artworks: creativity. We realise 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. 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 familiarise 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 design brief was given to the judges, and they were told this was the same brief seen by the participants, and that at the back was the instruction on technical support that was provided to all participants. Judges were then left to read these documents and the researcher set up the laptop to view the artwork. 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, see Figure 2 (developed to graphically reinforce the CAT protocol and instructions), 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 area designated "Medium", and judges proceeded to rate the artwork, and had as much time as they required.

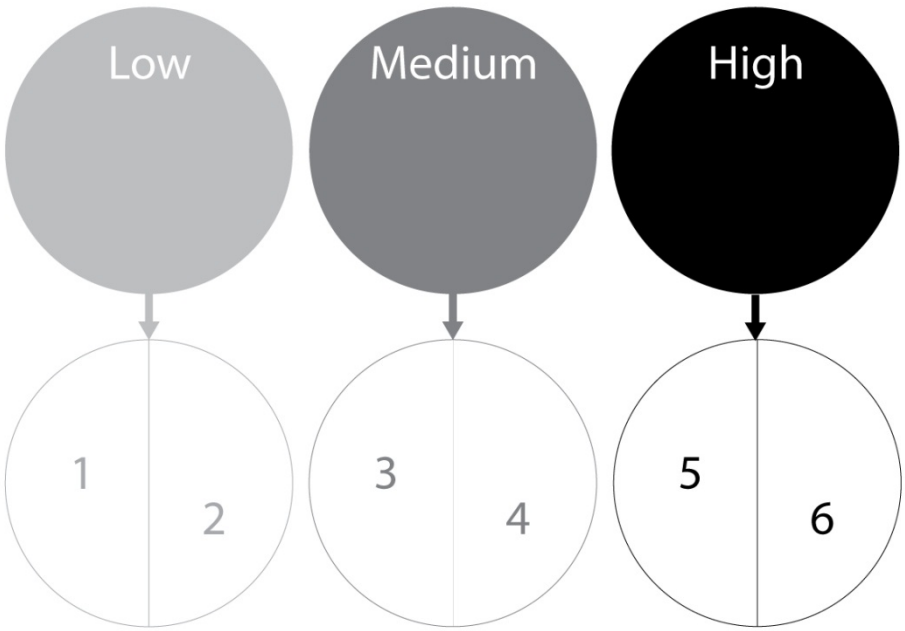

Figure 2. Rating sheet developed to graphically reinforce the CAT protocol

Once judges had read the instructions, three points were re-emphasized within the instructions. First, judges were to rate the artwork on creativity only (without considering technical execution and aesthetic appeal). This instruction is especially important to emphasise and draws a distinction between the CAT protocol in Jeffries (2015) with the present study. By verbally re-emphasizing this aspect of the instructions it could be argued that this additional procedure addresses whether a judge had read the

instructions sufficiently and whether the wording alone was robust enough to get the point across. Second, ratings were relative to the artworks provided, rather than an external standard of creativity (such as award-winning creativity). Third, it was important that the judges should divide the artworks more or less evenly over the ratings 1 to 6. The even number was to emphasise that judges use the whole of the 1 to 6 rating scale, and not overuse some of the scale (i.e. award everything a 1). A skew in rating can impact on the reliability of Cronbach's alpha, and would not be in keeping with the theoretical premise of the CAT being a relative measure of creativity within a defined sample.

Once the creativity assessment was completed, judges were given the instruction to rate either technical execution or aesthetic appeal. The previous creativity ratings were removed, the artworks were randomly re-sorted, then placed back to the top of the rating sheet. Judges were left to begin their assessment. With this completed, the same procedure was followed but for the remaining instructions, either aesthetic appeal or technical execution, which was determined by the randomization of judges into counterbalanced conditions. The wording of instruction for rating aesthetic appeal (see below) was virtually identical to that of technical executions: the only change was to replace the phrase aesthetic appeal with that of technical execution.

Aesthetic Appeal

Please look through these artworks and rate them for aesthetic appeal. 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 aesthetically appealing **(relative to the other artworks provided)**.

- Familiarise yourself with all the artworks provided.
- Group the artworks into Low, Medium, or High ratings.

- Assign a numerical rating between 1 and 6 (1's being the least aesthetically appealing and 6's being the most aesthetically appealing).

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.

4. Results

As discussed above, the expectation was that the inter-rater agreement for creativity would give a Cronbach's alpha above 0.7 and that there would be a moderate to strong significant correlation between technical execution, aesthetic appeal, and graphic design creativity.

The Inter-rater agreement for Creativity was a Cronbach's alpha of 0.92; for Aesthetic appeal, 0.88; Technical execution, 0.88. Each rating was above the 0.7 alpha level commonly used as a threshold for an appropriate standard of consensus among judges. The combined scores for each of the three ratings was computed and to check for normality of distribution; histograms were plotted for each rating. Skewness and standard error of skew were calculated to check if data would satisfy the assumptions of correlation coefficient analysis. All Z-skew statistics were within the threshold of +/- 1.96, suggesting the data was appropriate for parametric analysis. On this basis, Pearson's r was used to calculate the correlation coefficient between each rating. All correlations were positive at a significance level of 0.01 (1 tailed), and each suggestive of very strong correlations (Figure 3): creativity and aesthetic appeal (r: 0.93); aesthetic appeal and technical execution (r: 0.90); creativity and technical execution (r: 0.87).

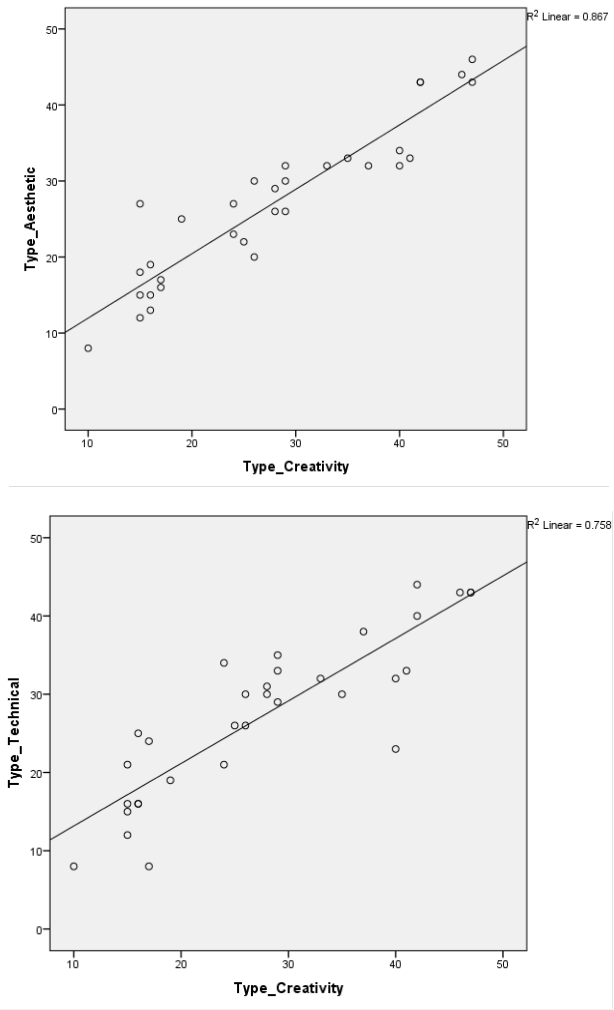

Figure 3. Scatter plots and correlations of creativity, technical execution, and aesthetic appeal.

In addition to the correlations above, a Principal Components Analysis (PCA) on the correlation coefficients was computed. The total variance, Scree plot, and component matrix are shown in Table 1 and 2 and Figure 4. Evidence can be found that a substantial single component is likely to be present which underpins judges' assessment of creativity, technical execution, and aesthetic appeal.

For the Type Task, only one principal component with an eigenvalue greater than one was identified, and the scree plot is characteristic of a data set with a single principal component. This single component explains 93% of the variance in the data.

Table 1: PCA of Type Task

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.803	93.430	93.430	2.803	93.430	93.430
2	.133	4.440	97.870	.133	4.440	97.870
3	.064	2.130	100.000	.064	2.130	100.000
Extraction Method: Principal Component Analysis.						

Table 2: Type Task PCA Component Matrix

Component Matrix[a]			
	Component		
	1	2	3
CREATIVITY	.966	-.212	.145
AESTHETIC	.977	-.073	-.199
TECHNICAL	.956	.288	.057
Extraction Method: Principal Component Analysis.			
a. 3 components extracted.			

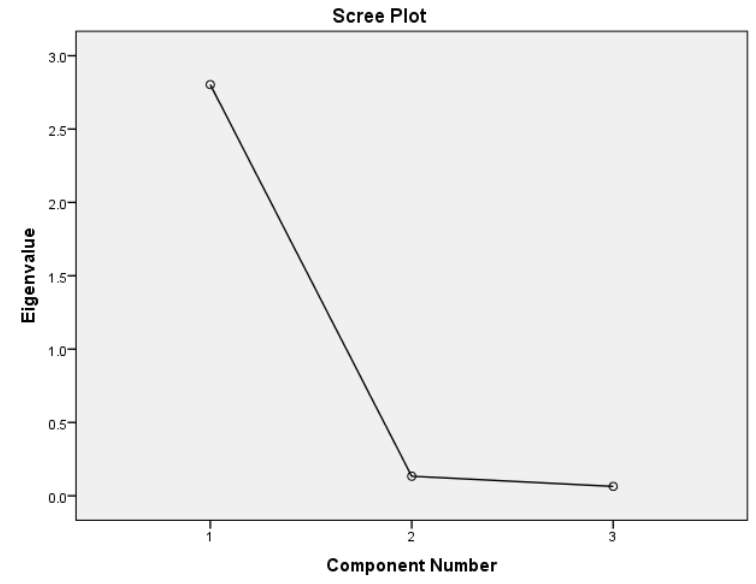

Figure 4: Scree plot for Type Task

5. Discussion

As highlighted in Jeffries, 2012, despite extensive use within creativity research, the use of the CAT as a measure of creativity within design research was relatively small.

Indeed, the study evidenced that only two design journal papers had operationalized CAT during a thirty-year period. Jeffries (2015) identified that the CAT could be a reliable measure of graphic design creativity, but consideration of a number of caveats was prudent, and the findings here suggest this has remained the case. Inter-rater agreement for creativity (alpha: 0.92) was considerably more than the 0.7 threshold, and a higher level of consensus than in Jeffries, 2015, (0.73), which ran the same tasks and had the same number of judges. This comparison, however, needs to be considered alongside the fact that judges were different, as was the artwork being judged. A further point was the population of participants in this study was more diverse than the 2015 study, and this may have impacted on the higher level of consensus.

Notwithstanding the limitations of a direct comparison between Jeffries (2015), and here, the increase in judges' consensus on creativity from 0.73 (in the 2015 study) to 0.92 (in this study) is likely to be significant. With more research using the Type Task, future experimental studies on CAT protocol will be able to identify the exact extent upon which this increase in alpha level is likely to occur. On the basis of published graphic design creativity studies, 0.92 is the highest known alpha to date.

Adjustments of CAT protocol were made in this study that may explain this increase beyond the influence of sampling bias/error. Notably, the verbal re-emphasis to judges not to compound their rating of creativity with technical execution or aesthetic appeal. If this is the case, then it lends further support towards the benefit of "creativity only" instructions, the value of verbal re-emphasis, and diversity of artwork. These caveats to CAT protocol do not seem to be detrimental to research design and would appear to improve the reliability, if not the validity, of CAT as a measure of creativity.

It is for this reason that the level of attention to procedure and precision of details regarding the CAT's application to graphic design creativity are present. For some, this degree of detail may be laboured (especially given such discussion is not always present in other scholars application of the technique), but the argument is that such details do matter. It is possible that creativity assessment in graphic design is particularly sensitive to such nuances of the method, and it may be that in other domains this degree of detail is not required. However, as is the case in other areas of design (like ergonomics) the value of designing for limiting users offers the opportunity to create a better "product" for all: the 95 percentile who fit within a population, as well as those found within the extremes.

As detailed in the method section, judges were asked first to assess creativity only, and then randomly assigned to judge technical execution or aesthetic appeal. A final assessment was either technical execution or aesthetic appeal depending on the previous outcome.

For some time, guidance has been that when a CAT task is developed for a different domain (i.e. one that has not been studied with a particular task), researchers should ask judges to rate both technical execution and aesthetic appeal, and check to see creativity ratings are distinct from these criteria. For the Type Task, each correlation was positive, significant and each suggestive of very strong correlations. It would appear that the Type Task is an obvious example of where creativity, technical execution, and aesthetic appeal are highly correlated.

In each of the scatter plots shown in the Results section (see Figure 3), and through the PCA, evidence can be found that a single substantial component is likely to be present that underpins the judges' assessment of creativity, technical execution, and aesthetic appeal. If this is the case for other design-focused CAT studies, then two key questions arise.

Firstly, is this original guideline on the distinction between creativity, technical execution and aesthetic appeal correct? Again, it must be acknowledged that Amabile and her colleagues made very clear this would not apply to some domains. Equally, the CAT was foremost used in social psychological research contexts, and showing discriminant validity between creativity and other compounding constructs was important to establish the validity of the measure at that time. Never the less, contradictory evidence in Amabile and Hennessey's work raises doubts about creativity and technical execution as distinct constructs. Within Amabile's 1982 studies she found correlations as high as .77 between creativity and technical goodness, and Hennessey (1994) presented statistically significant correlations as high as .71.

As identified in the Background section of this paper, other contemporary CAT research studies directly related to graphic design have identified high positive correlations (Wojtczuk & Bonnardel, 2012) between creativity and aesthetics. Equally, CAT inspired design studies exploring the creativity of product designs (Valgeirsdottir, Onarheim & Gabrielsen, 2015) have evidenced strong positive correlations between creativity and aesthetic appeal. For Industrial Design Engineering students, Christiaans' (2002) found strong correlations between creativity and aesthetic criteria, but not for creativity and technical quality.

In contrast, some design creativity researchers, when introducing a new CAT design task gather ratings for creativity, technical execution and aesthetic appeal but, given the focus of their study, do not report the correlations between these (Yuan & Lee, 2014); some design researchers using new CAT design tasks appear not to have gathered ratings from judges on technical execution or aesthetic appeal alongside creativity assessments. For example, it is unclear if Meneely and Portillo's 2005 study (with interior design students who undertook a furniture design task) had rated technical execution and aesthetic appeal in previous studies using this task. Given the range of interpretations of CAT protocols for design tasks, it would suggest the original guidance on evidencing a distinction between creativity, technical execution and aesthetic appeal is problematic for design creativity research.

Secondly, what does it mean if creativity, technical execution and aesthetic appeal are highly correlated? Moreover, what happens in those design domains where they are correlated: should researchers not use CAT? On one level it may be possible for detractors of CAT to see this evidence of confluence as a basis to say the method does not measure creativity. This is not the position of this paper: the argument here is that assessing creativity output is complex (especially graphic design creativity, and likely for other areas of design). Indeed, complexity is the reason why CAT did not work when first applied to graphic design. With various caveats considered, clearly, acceptable levels of consensus can be achieved. One of those caveats is how important it is to instruct judges not to consider technical execution and aesthetic appeal in their ratings of creativity. The case made here is further support for why the caveat on technical executions and aesthetic appeal is important. These constructs would appear to

be highly correlated for certain domains (graphic design for one), but this does not mean that judges cannot give a rating on creativity only; what is required is that this instruction is made clear to them. Emphasising that caveat enables a judge to tacitly reduce the background "noise" in creativity rating that is present from technical execution and aesthetic appeal. Isolating creativity may be a challenge, but the evidence is that they can achieve this. In contrast, if researchers do not do this, then every study could be skewing the creativity aspect with technical execution and aesthetic appeal. That has implications for comparison across studies.

It is worth highlighting that the CAT is considered to be a theoretically neutral measure of creativity, in that suitable judges are free to use whatever tacit or explicit criteria they may have evolved regarding creativity within their domain. In this respect, the CAT is quite distinct from other measures of creativity that define the criteria judges should apply when rating creativity: for example novelty, surprise, or usefulness. This theoretical neutrality is one reason creativity researchers have advocated the CAT as a “Gold Standard” of creativity assessment (Kaufman, Plucker & Baer, 2008).

The focus of this research study, and the previous study in 2015 is the influence of CAT instructions and protocols on the reliability of assessing graphic design creativity. The argument presented is that in domains (like graphic design) where the relationship between creativity, aesthetic appeal and technical execution are highly correlated, it is valuable for design researchers applying the CAT to use creativity only instructions and verbally re-emphasise these to maintain its theoretical neutrality. Without a researcher being explicit on creativity only ratings, the preferences of some judges for aesthetic appeal or technical executions would appear to impact on the reliability of their

creativity assessments. As an international design research community, our further exploration of the science of creativity now requires us to harmonise the tasks, instructions and protocols we used for the CAT. This study and the 2015 study are ultimately about what those CAT instructions and protocols may be for design creativity researchers.

5.1 Emergent considerations

One issue that emerged from the analysis was the level of detail for a graphic design task/brief. For example, the Type Task did not clarify which format the A4 sheet should be presented in; the majority of artworks from participants were landscape, while a few others chose portrait. To what extent such variation in format, or in some cases the layout of the graphic on the sheet, may influence the judges' consensus on creativity appears not to have had a direct impact overall on consensus, but it is likely that such choices could have had an influence on views of technical execution and/or aesthetics appeal.

From a design perspective, such choices by participants to opt for landscape or portrait was not problematic. The challenge came from the administration of the CAT. In the creation of a digital file to show judges, the artwork in portrait format was compounded by the default setting and limitations of PowerPoint. These constraints had not been apparent in the previous studies. Unfortunately, it was not straightforward to switch between landscape and portrait in a single PowerPoint file. For this reason, Adobe Illustrator was used instead, however, having created a PDF with both landscape and portrait artwork shown correctly, the challenge of randomising the artwork in Adobe Illustrator was more difficult. Artboards for each artwork in Adobe Illustrator needed to

be reordered by hand for each judge. This process appears not to be amenable to automation, and while laborious did achieve the PDF file showing a random order of artwork, and in the appropriate page orientation. Whether such additional procedures are worth the effort of offering participants the opportunity to select portrait and landscape formats is something to consider for future studies and discussion. For graphic design creativity the restriction to, for example, landscape only formats, for future research may be considered too restrictive a condition within the art and design community.

A final point on the issue of portrait or landscape format came from observing the judges as they undertook their assessments. Judges had four elements to engage with: an A4 hard copies of the design brief and instruction for rating the work; a laptop showing a PDF file that had all the artwork shown in the correct format (either landscape or portrait); an A3 laminated rating sheet; a set of laminated thumbnails of the artworks shown on the laptop that they could move around on the A3 rating sheet. What was noticeable, was that some artworks shown as portrait would occasionally find themselves in a landscape position on the rating sheet. Equally, as the rating progressed from creativity to technical execution, to aesthetic appeal, for example, judges tended to spend less time looking at the PDF file on the laptop and more time concentrating on the rating sheet. It was for this reason that each time after a judge had first assessed the creativity of the artworks, they were reminded that the details, both technical and aesthetic, would be better found on screen than in the thumbnails. It is perhaps a minor point given the high levels of consensus achieved in this study, but some artworks that were portrait may have suffered from being viewed as landscape thumbnails and judged more harshly. Especially, in those instances where the portrait format was crucial to

understanding the work, it is difficult to rule out they may not have been disadvantaged by the paper based rating sheet, and the dominance of predominantly portrait format artworks. Whether this is enough of a rationale to standardise the Type Task to a landscape only format, is difficult to say.

This issue may appear overly detailed, but it reflects a broader point regarding how standardized should a CAT graphic design creativity task be, or could become: did the current instructions give an opportunity for creativity because they did not specify a standard format to be followed by all participants, or does the lack of a standard format introduce more complexity and distract judges unnecessarily? It was possible to standardise all the artworks judges would view through opting for a landscape format. However, while this may enable easier administration and allow judges to view more easily all the artwork without the distraction of changes in format, for some artworks the choice of landscape or portrait was part of the quality of the work. The point highlights that while such consideration may not be of great importance in some domains, a graphic design CAT appears to offer a unique testing ground for design creativity researchers, one that is highly sensitive and responsive to the refinements of our methods.

6. Conclusion

Given the findings of this paper, it is worthwhile to reflect on the broader value of measuring technical execution and aesthetic appeal, alongside creativity. If assumptions of distinction are not met, but we have instructions and protocols that minimise the background "noise" of technical execution and aesthetic appeal on creativity assessment, then is rating additional constructs really worth the extra time and effort of

future design researchers who use the CAT? For example, Valgeirsdottir, Onarheim & Gabrielsen (2015) asked judges to rate three other constructs and creativity; the study above asked for two other constructs. In each study, rating additional components is a considerable amount of work when the reason for using the CAT is to assess creative output.

These are issues to be debated within the broader community of scholars. For the value of starting that debate, the conclusion of this paper is that if design creativity researchers want only to measure creativity, they should do so, and without this being viewed as a methodological flaw because of the original guidelines for the CAT. Obviously, such as statement is dependent on the research aims of a given study, but the arguments are that investment in a judge's time and effort, and the researcher's time and effort, make additional assessments unproductive. Recruitment to CAT studies is an acknowledged challenge as it currently stands; when the likelihood is there will be a highly significant and strong correlation between creativity, technical execution, and aesthetic appeal, it is questionable if that additional effort is worthwhile for the majority of future design creativity studies

Acknowledgments

removed

References

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

Hennessey, B. A., Amabile, T. M., & Mueller, J. S. (2011). Consensual assessment. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (2nd Ed., Vol. 1) (pp. 253-260). San Diego, CA: Academic Press.

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., & McKool, S. S. (2009). Assessing creativity using the Consensual Assessment Technique. In *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 1–13).

Chermayeff, I., & Geismar, T. H. (2006). *Watching Words Move*. Chronicle Books.

Christiaans, H. (2002). Creativity as a Design Criterion. *Creativity Research Journal,* 14, 41-54

Christiaans, H. & Venselaar, K. (2005). Creativity in design engineering and the role of knowledge: Modelling the expert. *International Journal of Technology and Design Education*, 15, 217-236.

Cseh, G. M. (2014). *Flow in visual creativity* (Unpublished doctoral dissertation). University of Aberdeen, Scotland.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction.* White Plains, NY: Longman.

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

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*), *1,* 211-220

Jeffries, K. K. (2015)**.** A CAT with caveats: Is the Consensual Assessment Technique a reliable measure of graphic design creativity? *The International Journal of Design Creativity and Innovation*.

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

Meneely, J., & Portillo, M. (2005). The adaptable mind in design: Relating personality, cognitive style, and creative performance. *Creativity Research Journal*, *17*(2-3), 155-166.

Valgeirsdottir, D., Onarheim, B., & Gabrielsen, G. (2015). Product creativity assessment of innovations: considering the creative process. *International Journal of Design Creativity and Innovation*, 3, 95-106.

Wojtczuk, A., & Bonnardel, N. (2012). Differences in creative design assessment. *Proceedings of the 2nd International Conference on Design Creativity (ICDC2012*).

Yuan, X., & Lee, J. H. (2014). A quantitative approach for assessment of creativity in product design. *Advanced Engineering Informatics*, *28*(4), 528-541.