Quasi-Experts' Assessments of Creative Products: An Evaluation Using a Sample of Design Students

Kuan Chen Tsai[1*]

[1]*Faculty of Humanities and Social Sciences, City University of Macau, China.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Original Research Article

ABSTRACT

The current study was designed to examine the possible use of quasi-experts (design-college students) to assess creative products (four-frame comic strips). The research questions we sought to answer were: (a) Can the consensual assessment technique (CAT) yield reliable ratings when quasi-experts are used as judges? And (b) can quasi-expert provides reliable ratings of comic strips across raters and across multiple comic strips? The results of our experiment using quasi-expert judges of visual creativity show significant positive correlations among the 10 rated comic strips across our five criteria of creativity. Another important finding is that the CAT is suitable for use with design students in a Chinese context. The current study indicated that art educators should encourage their art students to play the role of art critics, discussing the strengths and weaknesses of each other's work more frequently.

_____

*Corresponding author: E-mail: tsaikuanchen@cityu.edu.mo;

1. INTRODUCTION

Creativity has been studied from many perspectives, and has utilized various measurements aimed at capturing creative potential as exactly as possible [1]. One of the simplest methods is to have people self-assess their creativity and personality traits, using instruments such as the Kirton Adaption-Innovation Inventory (KAI) [2] or the Creative Personality Scale (CPS) [3]. However, as Kaufman, Lee, Baer, and Lee [4] remind us, "self-assessments are problematic because they are likely to include distortions of already suspect assessments" (p. 97). For this reason, Silvia, Wigert, Reiter-Palmon, and Kaufman [5] suggest that self-report scales may be more suitable for low-stakes assessment scenarios. Reiter-Palmon, Robinson-Morral, Kaufman, and Santo [6] have also recommended that researchers use caution when including self-perceptions of creativity as measures of creative performance.

Another popular battery of creativity tests, can be derived from Guilford's [7] theorization of creativity, attempts to measure cognitive skills: in particular, the so-called divergent-thinking abilities. Probably the most popular of these tests are the Torrance Tests of Creative Thinking (TTCT) [8], which score the phenomenon on four scales: fluency, flexibility, originality, and elaboration [9]. The validity of using divergent-thinking tests as a measure of creativity has been widely discussed in the creativity-research community [10-13], and although they have their detractors, on balance it is probably fair to say that divergent-thinking tests have a unique capacity to capture creative potential [14,15].

2. LITERATURE REVIEW OF ASSESSING CREATIVE PRODUCTS

Concerned by both the potential bias of self-assessments and the limitations of skill-measurements that are perhaps too closely rooted in specific theories of creativity, some scholars have pointed out that in real-world scenarios, it is common to ask a panel of experts in a given field to evaluate creative performance and products [16,17]: A procedure that Amabile [18] named the consensual assessment technique (CAT). Amabile [19] asked experts to evaluate three types of creative products – poems, stories, and collages – in ten dimensions, and then used factor analysis to extract three

factors: creativity, technical quality, and aesthetics. Many subsequent studies have found CAT to have acceptable inter-rater reliability of .70 to .90 [18,20,21]. It would therefore seem that CAT is a plausible approach to assessing creative performance via the collective wisdom of a group of experts. For most researchers, however, this time-consuming and costly technique is not practical. The present study therefore proposes and tests an alternative solution, similar to CAT but utilizing either quasi-experts or novices.

Besemer and Treffinger [22] proposed the Creative Product Analysis Matrix (CPAM) to help focus non-specialist judges' attention on the most relevant attributes of creative products. Besemer and O'Quin [23] later refined this model and developed it into the Creative Product Semantic Scale (CPSS), a bipolar instrument analyzing a product's creativity in terms of three dimensions: novelty, resolution, and elaboration/synthesis. This three-dimensional schema, which its authors argued was easy to grasp and a helpful tool for creators undergoing self-training, is further divided into a total of 14 subscales. These are: *germinal, original, transformational* (for novelty); *adequate, appropriate, logical, useful*, and *valuable* (for resolution); and *attractive, complex, elegant, expressive, organic*, and *well-crafted* (for elaboration/synthesis). The validity of the CPSS was confirmed by Besemer and O'Quin's [24].

After reviewing the definitions of creativity used in 90 empirical studies across different fields, Dean, Hender, Rodgers, and Santanen [25] proposed that it has four dimensions: novelty, relevance, workability, and specificity, a framework originally proposed by MacCrimmon and Wagner [26]. In other words, these authors define a *creative idea* as a high-quality idea with the novel attribute that is an effective and workable solution to a problem. In the same paper, Dean et al. went on to develop a multidimensional construct for measuring the creativity of ideas. This subdivided novelty into *rarity, originality*, and *paradigm -relatedness*; relevance into *applicability* and *effectiveness*; workability into *implementability* and *acceptability*; and specificity into *clarity, completeness*, and *implicational explicitness* (for more detail, please see pp. 658-662). Following principal component analysis and confirmatory factor analysis, clarity was dropped from the dimension of specificity, which left nine sub-scales in the construct.

A number of empirical studies have attempted to identify proper candidates for the assessment of creative products. Christiaans [27] reported on two such experiments. In the first, 44 designs were submitted for judgment by experts (ten design teachers), novices (12 mathematics students), and quasi-experts (12 design students), who assessed the work for creativity, prototypical value, attractiveness, interest, technical quality, expressiveness, and integrating capacity. The results indicated that the reliability of creativity ratings across the three groups of raters was acceptable. Moreover, experts' judgments were found to be no more consistent than those of novices; and of the three groups, the quasi-experts were actually most able to differentiate between the assessment attributes. In Christiaans' second experiment, 10 quasi-experts were selected to judge 55 design products using CPSS. Again, a high level of agreement was found. Christiaans therefore suggested that the selection of a homogenous group of judges would increase the reliability and validity of judgments. Another study involving quasi-experts [28] presented three-dimensional artworks by 47 art students for judging their creativity by three panels, comprising (a) their own creators, (b) other art students, and (c) professional artists. The results suggest that professional artists may not be the best option of evaluating nonprofessional artwork; rather, peer and self-rating might be more tenable for investigating the creative potential of students.

Kaufman et al. [4] asked four graduate students (three in psychology and one in history) to serve as novice raters of 12 captions produced by 81 undergraduates, using CAT. The study found a high level of inter-rater reliability. Kaufman, Baer, and Cole [29] compared the judgments of novices (106 college students) and experts (10 writers) on 203 pieces of student short fiction, and found that the novices' creativity ratings were quite different than those of the experts, with the experts being far more consistent in their ratings. Another study comparing experts and novices [30] also found that experts and non-experts had differential levels of inter-rater agreement, suggesting that expert judges agreed to an appropriate degree, whereas non-expert judges did not. The authors of that study urged caution in the use of non-expert raters, and called for future research to compare the judgments produced by experts, quasi-experts, gifted novices, and novices. Following this suggestion, Kaufman, Gentile, and Baer [31] asked gifted novices (high-school creative writers)

and three groups of experts (cognitive psychologists, professional creative writers, and high-school teachers) to rate the creativity of 27 short stories and 28 poems written by other groups. The inter-rater agreement within and among the three groups of experts was strong, but the agreement within the group of gifted novices was also acceptable, leading support to the use of peer feedback among gifted novice creative writers.

As the preceding discussion has implied, most studies of this type conducted on data have focused on novice- and expert judging verbal creativity, whereas little research has been conducted on quasi-experts' judgments of figural creative products. Therefore, the current study was designed to examine the possible use of quasi-experts (design-college students) to assess creative products (four-frame comic strips). The research questions we sought to answer were: (a) Can the CAT yield reliable ratings when quasi-experts are used as judges? And (b) can quasi-expert provide reliable ratings of comic strips across raters and across multiple comic strips?

3. METHODS

3.1 Participants

To test the use of quasi-experts for evaluating visual creative products, 84 second-year Chinese undergraduates taking graphic design courses at a small university in Macau were recruited as raters. The 32 men and 52 women, with an average age of 19 years, were given extra credit in return for their participation.

3.2 Materials and Procedures

We collected 10 four-frame comic strips from a Chinese comics website (http://comic.zongheng.com/series/c2.html) to which anyone can upload his comic creation to be viewed by others free of charge. In light of the preceding discussion of the relevant literature, the current study employed five dimensions – originality, quality, attractiveness, technical quality, and elegance – as criteria for the students to evaluate these comic strips. Data collection was conducted in a campus computer laboratory, where the participants were given instructions for the task. Each comic strip was viewed on the computer screen and rated using the five above-mentioned dimensions, on a scale

ranging from 1 (lowest) to 5 (highest). The whole process took less than 20 minutes to complete. All raters worked independently and received identical instructions.

4. RESULTS

A series of Pearson correlations was conducted among our raters and the five dimensions of their ratings. The results revealed significant relationships among the ratings of the 10 comic strips in the dimension of originality, ranging from $r = .22$ to $.68$, $p < .01$. For quality, significant relationships were also found, from $r = .30$ to $.70$, $p < .01$. For attractiveness, the majority of the strips exhibited significant relationships, from $r = .23$ to $.64$, $p < .05$, but the correlation of ratings between comic strips numbers 2 and 6 was low, at $r = .16$, $p = .14$. For technical quality, significant relationships were mostly found, ranging from $r = .24$ to $.82$, $p < .05$, with the exception against strips 2 and 6 ($r = .08$, $p = .45$). In the dimension of elegance, similarly, significant relationships (from $r = .24$ to $.65$, $p < .05$) were found except the case of strips 6 and 8 ($r = .20$, $p = .07$). Overall, the correlations among ratings on the five dimensions were significant, with effect sizes ranging from weak to strong.

Inter-rater reliability was assessed via Cronbach's alpha as shown in Table 1, and was found to be in the "excellent" range for all five dimensions of creativity that our panel assessed. Specifically, the coefficient alpha of originality was .891 (95% CI = .852 - .922; $M = 3.63$, variance = .05). For quality, it was also .891 (95% CI = .853 - .923; $M = 3.37$, variance = .10).

For attractiveness, it was .859 (95% CI = .809 - .900; $M = 3.25$, variance = .09); for technical quality, .889 (95% CI = .850 - .921; $M = 3.32$, variance = .11); and for elegance, .883 (95% CI = .842 - .917; $M = 3.20$, variance = .10).

The indices of originality, quality, attractiveness, technical quality, and elegance were calculated using the average scores of the 10 comic strips in each of our five dimensions. Table 2 shows that intercorrelations among these five dimensions and all correlations were significant ($p < .01$) and strong (from .80 to .97).

Varimax-rotated principal component analysis was also used. The result of a Bartlett's test of sphericity was statistically significant ($p < .001$), indicating that sufficient correlations exist among the variables; a one-factor solution was also found, accounting for 90.12% of the variance. The commonalities of the variables ranged from .828 to .961, and their factor loadings from .910 to .980. The single factor can be labeled as creativity, and incorporates our five dimensions of creativity as its sub-components.

5. DISCUSSION

The results of our experiment using quasi-expert judges of visual creativity (comic strips) show significant positive correlations among the 10 rated comic strips across our five criteria of creativity, ranging from $r = .22$ to $.82$. In terms of inter-rater reliability, coefficient alpha attained a satisfactory level, from .859 to .889. Both results indicate that design students can effectively serve as panels of judges to assess creative products. Among the five dimensions of creativity

Table 1. Means, variances and Cronbach's alpha for scores on five measures of comic strips

Measure	Means	Variances	Cronbach's alpha	95% CI
Originality	3.63	.05	.891	.852 - .922
Quality	3.37	.10	.891	.853 - .923
Attractiveness	3.25	.09	.859	.809 - .900
Technical quality	3.32	.11	.889	.850 - .921
Elegance	3.20	.10	.883	.842 - .917

Table 2. Intercorrelations for scores on five measures of comic strips

Measure	1	2	3	4	5
1. Originality	--				
2. Quality	.87**	--			
3. Attractiveness	.83**	.89**	--		
4. Technical quality	.82**	.97**	.86**	--	
5. Elegance	.80**	.92**	.88**	.91**	--

*** p < .01*

we used, originality and quality obtained the most stable agreement among our quasi-expert judges, whereas attractiveness was the quality they agreed upon as the least. These findings indicate that CAT can yield reliable ratings when quasi-experts are used as judges and our Chinese quasi-experts can also provide reliable ratings of comic strips across raters and across multiple comic strips.

Another important finding is that the CAT is suitable for use with design students in a Chinese context, confirming that the validity of the CAT is grounded in the expertise of the judges and not in cultural or other factors. The findings of the current study also suggest that the CAT can be used with quasi-expert. Overall, the results support previous studies finding on the usefulness of the CAT [20,30].

The current study also carries an important implication for educators. Peer feedback, which can enhance collaborative learning, is an important teaching strategy in consensual assessment [31]. To the extent that it has confirmed the reliability of quasi-expert judgment. This research therefore lends support to the use of peer feedback as an important resource, not only for learning by its content, but also as a tool for learning how to critique creative works. As such, art educators should encourage their art students to play the role of art critics, discussing the strengths and weaknesses of each other's work more frequently.

For researchers, the major drawback of using CAT in creativity research has been that it is time-consuming and often costly. However, as compared to recruiting a panel of experts, quasi-expert judges are likely to be easier to approach, not least because they may be self-organized into coherent groups (e.g., college classes) in a way that well-known professional practitioners are not. Nevertheless, we recommend that future studies can examine the use of quasi-experts across multiple domains. However, whether similar findings to ours would be found in the visual arts, performing arts, and other such fields remains unknown. Our study provides an important contribution toward knowledge in this area, but it is also clear that more research is needed before quasi-expert CAT can be recommended for general adoption.

In the views of Kaufman and Baer [32], "whether researchers use the CAT strictly with appropriate experts or whether they are more lenient and use trained novices, quasi-experts, or expert-less domains, … it is incumbent upon researchers to briefly discuss these issues" (p. 90). If, as our findings suggest, it is usually appropriate to use quasi-expert judges to assess creative performance, this has important practical implications for other similar studies, especially in terms of reducing the burdens of time and cost that have hitherto been associated with recruiting panels of experts.

6. LIMITS

Before turning to the broader implications of this study, limitations should be mentioned. First, the study was conducted with a homogeneous selection of students from a single institution and with the same cultural background. Future work will need to include multiple learning sites, and participants with multicultural heritage. In addition, the comic strips were used in the present study, and in future research other forms of creative products (e.g., drawing and sculpture) should be examined to further validity the results.

7. CONCLUSION

In our study we asked undergraduates in Macau to assess five dimensions of creative products—originality, quality, attractiveness, technical quality, and elegance—and found that intercorrelations among ratings on the five dimensions were significant and the coefficient alpha of inter-rater reliability was also good. These findings indicate that the CAT yield reliable ratings by using quasi-expert judges and Chinese college students serving as quasi-experts provide reliable ratings of comic strips across raters and across multiple comic strips. In short, our findings suggest that the potential use of quasi-experts as a panel of judges for visual creativity.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Hocevar D, Bachelor P. A taxonomy and critique of measurements used in the study of creativity. In: Glover JA, Ronning, RR, Reynolds CR, editors, Handbook of creativity. New York: Plenum Press; 1989.
2. Kirton MJ. Adaptors and innovators: A description and measure. Journal of Applied Psychology. 1976;61(1):622-29.

3. Gough HG. A creative personality scale for the adjective check list. Journal of Personality and Social Psychology. 1979;37(8):1398-1405.

4. Kaufman JC, Lee J, Baer J, Lee S. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. Thinking Skills and Creativity. 2007;2(1):96-106.

5. Silvia PJ, Wigert B, Reiter-Palmon R, Kaufman JC. Assessing creativity with self-report scales: A review and empirical evaluation. Psychology of Aesthetics, Creativity and the Arts. 2012;6(1):19-34.

6. Reiter-Palmon R, Robinson-Morral EJ, Kaufman JC, Santo JB. Evaluation of self-perceptions of creativity: Is it a useful criterion? Creativity Research Journal. 2012;24(2-3):107-14.

7. Guilford JP. Creativity. American Psychologist. 1950;5(9):444-54.

8. Torrance EP. Torrance Tests of Creative Thinking. Personnel Press; 1974.

9. Kim KH. Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). Creativity Research Journal. 2006;18(1):3-14.

10. Feldhusen JF, Goh BE. Assessing and accessing creativity: An integrative review of theory, research, and development. Creativity Research Journal. 1995;8(3):231-47.

11. Hocevar D. Measurement of creativity: Review and critique. Journal of Personality Assessment. 1981;45(5):450-64.

12. Plucker JA. Is the proof in the pudding? Reanalyses of Torrance's (1958 to present) longitudinal data. Creativity Research Journal. 1999;12(2):103-14.

13. Torrance EP. Predictive validity of the torrance tests of creative thinking. The Journal of Creative Behavior. 1972;6(4):236-62.

14. Cropley AJ. Defining and measuring creativity: Are creativity tests worth using? Roeper Review. 2000;23(2):72-9.

15. Runco MA, Millar G, Acar S, Cramond B. Torrance test of creative thinking as predictors of personal and public achievement: A fifty-year follow-up. Creativity Research Journal. 2010;22(4):361-68.

16. Hennessey BA, Amabile TM. Consensual assessment. In: Runco MA, Pritzker SR, editors. Encyclopedia of creativity. San Diego: Academic Press; 1999.

17. Kaufman JC, Baer J, Agars MD, Loomis D. Creativity stereotypes and the consensual assessment technique. Creativity Research Journal. 2010;22(2):200-05.

18. Amabile TM. Creativity In context: Update to the social psychology of creativity. Boulder: Westview Press; 1996.

19. Amabile TM. Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology. 1982;43(5):997-1013.

20. Baer J, Kaufman JC, Gentile, CA. Extension of the consensual assessment technique to nonparallel creative products. Creativity Research Journal. 2004;16(1):113-17.

21. Baer J, McKool SS. Assessing creativity using the consensual assessment technique. In: Schreiner CS, editor. Handbook of research on assessment technologies, methods, and applications in higher education. Hershey: IGI Global; 2009.

22. Besemer SP, Treffinger DJ. Analysis of creative products: Review and synthesis. Journal of Creative Behavior. 1981;15(3):158-78.

23. Besemer SP, O'Quin K. Analyzing creative products: Refinement and test of a judging instrument. The Journal of Creative Behavior. 1986;20(2):115-26.

24. Besemer SP, O'Quin K. Confirming the three-factor creative product analysis matrix model in an American sample. Creativity Research Journal. 1999;12(4):287-96.

25. Dean DL, Hender JM, Rodgers TL, Santanen EL. Identifying quality, novel, and creative ideas: Constructs and scales for idea evaluation. Journal of the Association for Information Systems. 2006;*7*(10):646-99.

26. MacCrimmon KR, Wagner C. Stimulating ideas through creativity software. Management Science. 1994;40(11):1514-532.

27. Christiaans HHCM. Creativity as a design criterion. Creativity Research Journal. 2002;14(1):41-54.

28. Runco MA, McCarthy KA, Svenson E. Judgments of the creativity of artwork from students and professional artists. The Journal of Psychology. 1994;128(1):23-31.

29. Kaufman JC, Baer J, Cole JC, Sexton JD. A comparison of expert and nonexpert raters using the consensual assessment technique. Creativity Research Journal. 2008;20(2):171-78.

30. Kaufman JC, Baer J, Cole JC. Expertise, domains, and the consensual assessment technique. Journal of Creative Behavior. 2009;43(4):223-33.

31. Kaufman JC, Gentile CA, Baer J. Do gifted student writers and creative writing experts rate creativity the same way? Gifted Child Quarterly. 2005;49(3):260-70.

32. Kaufman JC, Baer J. Beyond new and appropriate: Who decides what is creative? Creativity Research Journal. 2012;24(1): 83-91.