Creativity Research Journal

ISSN: 1040-0419 (Print) 1532-6934 (Online) Journal homepage: www.tandfonline.com/journals/hcrj20

Does the Picture Concept Task Measure Divergent and Convergent Creativity? A Validation Study

Weitao Zhang, Luisa Prochazkova & Bernhard Hommel

To cite this article: Weitao Zhang, Luisa Prochazkova & Bernhard Hommel (2026) Does the Picture Concept Task Measure Divergent and Convergent Creativity? A Validation Study, Creativity Research Journal, 38:3, 528-536, DOI: 10.1080/10400419.2025.2478728

To link to this article: https://doi.org/10.1080/10400419.2025.2478728

CREATIVITY RESEARCH JOURNAL
2026, VOL. 38, NO. 3, 528–536
https://doi.org/10.1080/10400419.2025.2478728

Routledge
Taylor & Francis Group

Does the Picture Concept Task Measure Divergent and Convergent Creativity? A Validation Study

Weitao Zhang[a,b], Luisa Prochazkova[b], and Bernhard Hommel[c]

[a]Zhejiang University; [b]Leiden University; [c]Shandong Normal University

ABSTRACT

The Picture Concept Task (PCT) has been frequently used in previous studies to measure convergent thinking (CT) and divergent thinking (DT) of creativity (the sub-tests for CT and DT are referred to as PCTc and PCTd, respectively), but it has not yet been validated. Therefore, this study aimed to validate the PCT using Classical Test Theory with a sample size of 158 participants, specifically by performing item analysis to assess difficulty and discrimination, and by calculating Cronbach's alpha to evaluate reliability. Additionally, the validity was evaluated by calculating the correlations between the PCT and the typical CT and DT tasks. The results indicate that, overall, both the PCTc and PCTd are inadequate for measuring CT and DT, as PCTc and two scoring dimensions of PCTd show poor performance in item analysis, reliability, and validity. However, the scoring dimension of fluency in PCTd can be a good indicator of the fluency of DT, with both reliability and validity yielding satisfactory results. Nevertheless, with improvements to the scoring method for the dimension of originality, the PCTd may prove to be a valid measure of DT.

ARTICLE HISTORY
Received November 17, 2023

Introduction

Creativity is a complex and multidimensional phenomenon, and since the inception of creativity research, creativity researchers have proposed numerous definitions based on various dimensions (Cropley, 2020). When considering creativity from the perspective of creative outcomes or products, the definition generally requires both originality and effectiveness (Runco and Jaeger, 2012). In terms of the cognitive processes involved in creativity, Guilford (1967) viewed creativity as a form of problem-solving and distinguished between two types of cognitive operations: divergent thinking and convergent thinking. Mednick (1962) suggested that creativity involves combining associative elements in novel ways to meet task demands. From the perspective of creative ability, creativity is thought to require both intelligence and expertise (Stevenson et al., 2021). Regarding personality, several personal characteristics and traits have been linked to creativity, including attraction to complexity, high energy, behavioral flexibility, intuition, emotional variability, self-esteem, risk-taking, perseverance, independence, introversion, social poise, and tolerance for ambiguity (Said-Metwaly et al., 2017). Based on these various dimensions and definitions, researchers have developed different tests and scoring methods. For instance, in line with the definitions of divergent and convergent thinking, the Alternative Uses Task (AUT) and Remote Associates Test (RAT) can be used to measure divergent and convergent thinking, respectively (Zhang et al., 2020). Additionally, RAT is also used to assess associative processes (Mednick, 1968). To evaluate originality and effectiveness, researchers often employ the Consensual Assessment Technique (CAT; Baer & Kaufman, 2019), while the Self-Report of Creative Traits (Runco et al., 2017) is used to assess creativity-related personality traits. As definitions of creativity evolve and research contexts shift, new assessments continue to be developed to better capture the multifaceted nature of creativity.

In recent years, the Picture Concept Task (PCT), a picture-based test, has been increasingly used to measure creativity, particularly in research on psychedelics (Girn et al., 2020; Kuypers et al., 2016; Mason et al., 2019; Prochazkova et al., 2018). The PCT consists of matrices of pictures (e.g., 2 × 3, 3 × 3, 3 × 4; see Figure 1) selected from the Wechsler Intelligence Scale for Children (WISC; Kuypers et al., 2016; Wechsler, 2003), and participants are tasked with finding common associations among the presented pictures (see Figure 1). Based on Guilford's (1967) definition of creative thinking, the PCT has been developed into two subversions to measure convergent thinking (CT) and divergent

CONTACT Bernhard Hommel bh@bhommel.onmicrosoft.com School of Psychology, Shandong Normal University, No. 88 East Wenhua Road, Jinan, Shandong 250014, China

Figure 1. Example item from the picture concept task. In the PCTc, participants are asked to select one picture from each row and provide the common association (e.g., choose the picture 3, 5, and 7 with an answer "water"). In the PCTd, participants are asked to choose any two pictures from any two different rows and generate as many common associations as possible (e.g., selecting the picture 6 and 9 with an answer "tree;" selecting the picture 1 and 8 and describing their common shape as the common association; . . .).

thinking (DT). The convergent thinking version (PCTc) requires participants to find the only correct common association, with evaluation based on response accuracy. The divergent thinking version (PCTd) asks participants to find as many common associations as possible, and it is evaluated across three dimensions: fluency (the number of alternative associations provided), originality (an answer earns 2 points if reported by only one participant, 1 point if reported by two participants, and 0 if more), and originality-to-fluency ratio (originality divided by fluency).

Although the PCT has been used in many studies (e.g., Kuypers et al., 2016; Mason et al., 2019, 2021; Prochazkova et al., 2018; Uthaug et al., 2019; Wießner et al., 2022), a standardized and properly validated version does not appear to exist. As a result, it is unclear whether the PCT qualifies as a reliable test for creativity. Notably, to the best of our knowledge, the earliest study using the PCT to assess creativity was conducted by Kuypers et al. (2016), in which the PCT was derived from the WISC and categorized under the Fluid Reasoning Index (Wechsler, 2003). While the two sub-tests, PCTc and PCTd, seemingly align with the definitions of CT and DT, it is important to recognize that the PCT was originally designed to assess categorical and abstract reasoning capacity (Wechsler, 2003). Additionally, in the WISC, the task requirements of the Picture Concepts Task (PCT) are more similar to PCTc than PCTd. According to the WISC-IV manual, the PCT used in the WISC has good reliability, with a test–retest reliability of $r = .71$ and an internal consistency of $r = .83$ (Wechsler, 2003). It also serves as a valid measure of fluid reasoning, showing a stronger correlation with Matrix Reasoning ($r = .47$) than with other WISC-IV subtests (Sattler & Dumont, 2008). Furthermore, it has a moderate-to-high correlation with the Perceptual Reasoning Index ($r = .77$) and a moderate correlation with the Verbal Comprehension Index ($r = .47$), suggesting that Picture Concepts is a better indicator of perceptual reasoning than verbal abilities (Sattler & Dumont, 2008).

However, the validity of the PCT as a measure of creativity remains unclear. Since the PCT has primarily been used in psychedelic research, its application and statistical analysis are typically conducted by comparing pre- and post-psychedelic use, using ANOVA or t-tests. However, neither the initial study by Kuypers et al. (2016) nor subsequent studies following its methodology have provided information regarding the test's reliability or validity, or its correlation with other creativity measures (e.g., Mason et al., 2019, 2021; Prochazkova et al., 2018; Uthaug et al., 2019; Wießner et al., 2022). Moreover, the scoring method used in the PCTd, particularly in previous studies, also gives rise to some concerns. The frequency-based originality score is highly influenced by sample size – the larger the sample, the less likely a response is to score highly. This method has also been criticized for potentially awarding random,

vague, or nonsensical responses (Forthmann et al., 2021; Silvia et al., 2008). Additionally, the reliability of the originality-to-fluency ratio is questionable due to the technical limitations of ratio scores (Cronbach, 1941) and is further impacted by variability in originality scores. Another issue is that the instructions for the PCT do not explicitly emphasize creativity but merely ask participants to find the common association between the pictures, which is a necessary condition for a reliable and valid creativity assessment (Acar et al., 2020). These points raise legitimate concerns about the appropriateness of the PCT as a tool for assessing creativity.

The main research question of the present study was to examine whether the PCT is a valid instrument for assessing creativity. There are two common methods for validating a test: Classical Test Theory (CTT) and Item Response Theory (IRT). IRT models the probability of a correct response based on item difficulty and participant ability, while CTT focuses on observed test scores and their reliability across repeated measurements (Jabrayilov et al., 2016). It is important to note that IRT typically demands a much larger sample size than CTT (Zanon et al., 2016). As our resources did not suffice to test such a large sample, we chose to employ CTT to assess the PCT. We believe this approach is adequate for addressing our research question, which revolves around evaluating the quality of the PCT, with the suspicion that it may not perform favorably. In addition, nevertheless, we also conducted an IRT analysis to complement and support the analysis conducted using CTT, and provided the results on Open Science Framework.

In the CTT analysis of this study, specifically, we conducted the item analysis to assess difficulty and discrimination, and calculated the internal consistency reliability using Cronbach's alpha. In terms of validity analysis, since the PCTd and PCTc are designed to measure DT and CT, the validity (both convergent and discriminant) of the PCT was analyzed by correlating it with typical DT and CT tests. However, it is important to note that while the AUT and RAT are classical measures of these constructs, both tasks actually involve processes related to both DT and CT (Cortes et al., 2019). Although the strict dichotomy between CT and DT is questionable (Gabora, 2018), emphasizing this distinction remains valid in test design and measurement so that the use of AUT and RAT remains a relatively reliable choice. Since CT and DT are conceptualized as opposing constructs, a well-validated PCTc should exhibit a positive correlation with RAT and a negative or no correlation with AUT. Conversely, the well-validated PCTd should display a positive correlation with AUT, and a negative or no correlation with RAT.

Methods

Participants

We tested 158 participants (151 females and 7 males) in this study. All of them were Dutch native speakers and were studying at Leiden University, the Netherlands. Their ages ranged from 18 to 32 years with a mean of 19.0 (SD = 1.9). They were tested for 90 minutes online using Qualtrics software (Qualtrics, Provo, UT). The study was approved by the ethical committee for Leiden University (Faculty of Social and Behavioral Sciences) and informed consent was given by the participants.

Instruments

Picture concept task

The PCT (e.g., Kuypers et al., 2016; Prochazkova et al., 2018; Wechsler, 2003) is a picture-based creativity task that requires participants to find common association(s) from a group of pictures. Each item in it consists of a matrix of between 2×3 to 3×4 pictures (Figure 1 as an example), and participants are instructed to select one picture from each row to find common association(s). The PCT includes two sub-tests (PCTc and PCTd) that assess CT and DT, respectively. There were 32 items in the PCTc and 16 items in the PCTd, obtained from a previous study (Prochazkova et al., 2018), for validation in this study (the materials can be found at: https://osf.io/yxh6s/).

In the PCTc, there is only one correct solution for each item, and participants are asked to find the common association in 35 seconds. The score of PCTc for each participant is calculated by summing the number of correct answers.

In the PCTd, participants are asked to find as many associations as possible within 2 minutes, and each association should be formed between any two different pictures from any two different rows. Scoring for the PCTd involves three dimensions: fluency, originality, and the ratio of both (originality divided by fluency). Fluency is determined by the number of alternative associations provided. In most of the previous studies that used the PCT (Kuypers et al., 2016; Mason et al., 2019, 2021; Uthaug et al., 2019), the originality is based on the number of participants who share the same answer (if the answer is reported by only one participant, the originality score is 2; if reported by two participants, the score is 1; otherwise, it is 0). To reduce the

impact of the fluency confounding problem, we calculated the average originality score for the response provided. The ratio is calculated by dividing the mean originality score by the mean fluency score.

The Qualtrics software (Qualtrics, Provo, UT) was used for presenting the task and recording the answers. Participants can click on images using a mouse and then enter their answers in a textbox.

Remote associates task

The Remote Associates Task (RAT) is a typical task measuring CT (Guilford, 1967). In the RAT, three unrelated words are presented (e.g., time, hair, and stretch), and participants are required to generate a fourth word (the answer is "long" in this example) which can connect with the three given words. In this study, a Dutch-version compound RAT (Akbari Chermahini et al., 2012) consisting of 30 items was used for the validity analysis of the PCT. This RAT shows good reliability (Cronbach's alpha = 0.85) and convergent validity revealed by its relation to Raven's Advanced Progressive Matrices test and insight-problem test and discriminant validity revealed by its relation to Guilford's Alternative Uses Test. For each item, participants were given 35 seconds to think and enter the answer in a textbox. The score of RAT was calculated by summing the number of correct answers.

Alternate uses task

The Alternate Uses Task (AUT) is a typical task to assess DT (Guilford, 1967), which was used for the validity analysis of the PCT. In the AUT, a common object (e.g., cup) is presented and participants are asked to generate as many creative uses of the object as possible (e.g., flowerpot, make art, fishbowl, . . .). In the instructions, both "creative" and "as many as" were emphasized. In this study, participants were given two items (fork and shoes) and had 2 minutes for each item to enter their ideas in textboxes. The scores of AUT were evaluated based on three dimensions: fluency, flexibility, and originality. Fluency is simply scored based on the number of responses provided. To score flexibility and originality, two bachelor students were trained as raters. For flexibility, these two raters first reviewed all the responses and collaboratively developed a common list of categories. One rater then assigned the responses to the appropriate categories, and the total number of distinct categories was used to calculate the flexibility score. For originality, the two trained raters independently scored the originality of each idea from each item on a scale of 1–5, representing the original level. Inter-rater reliability was calculated using a two-way mix intraclass correlation coefficient analysis for consistency which showed a good result (intraclass correlation coefficient = 0.76). During scoring, duplicate responses were counted as a single response, while incomplete and nonsensical responses were excluded. To reduce the impact of fluency confounding problem, the originality score for each item is the average originality score of all responses, and each participant's originality score is the average of the originality scores across all items. The final originality score was obtained as the mean score of the two raters. To evaluate the impact of the fluency confounding problem, a delta test was applied (Forthmann et al., 2020) to AUT-fle and AUT-ori. In both tests, the significance level of delta-related correlations was $p < .001$, meaning that the correlation between fluency and originality is not solely due to fluency confounding, and the raw (observed) correlation largely reflects true originality.

Data analysis

The Classical Test Theory (CTT) was applied to analyze the PCT and its items using IBM SPSS statistics (Version 27.0). First, item analysis was conducted, and the probability of valid solutions for PCTc and the mean score for PCTd were used as indicators of difficulty, and corrected item-total correlations were calculated to evaluate discrimination. As the item of PCTd is not a dichotomous question, the standard deviation, skewness, and kurtosis of its responses were also provided to understand the distribution of the scores. Additionally, an item-total correlation higher than 0.3 (Cristobal et al., 2007) was adopted as an acceptable criterion. Secondly, reliability was evaluated by Cronbach's alpha, which is considered a good criterion for measuring internal consistency reliability when it exceeds 0.7 (Taber, 2018). Finally, it is important to consider convergent and discriminant validity in assessing the PCT. The PCTc, which is designed to measure CT, is expected to demonstrate a positive correlation with the RAT. Similarly, the PCTd, aimed at measuring DT, should exhibit a positive correlation with DT. On the other hand, a good discriminant validity would be demonstrated by negative or non-significant correlations between the PCTc and AUT, as well as between the PCTd and RAT.

The mentioned analysis based on Item Response Theory (IRT) is available on Open Science Framework (https://osf.io/t6ekp/).

Table 1. The results of the item analysis for PCTc.

Item	Probability of valid solution	Item-total correlation	Item	Probability of valid solution	Item-total correlation
1	0.62	0.18	17	0.88	0.24
2	0.57	0.11	18	0.14	−0.02
3	0.48	0	19	0.04	−0.04
4	0.38	0.08	20	0.02	0.09
5	0.44	0.19	21	0.06	0.11
6	0.76	0.21	22	0.41	0.16
7	0.5	0.16	23	0.53	0.03
8	0.59	0.13	24	0.33	0.11
9	0.37	0.04	25	0.55	0.06
10	0.55	0.23	26	0.23	−0.2
11	0.57	0	27	0.25	−0.01
12	0.79	0.08	28	0.34	0.17
13	0.59	0.09	29	0.29	0.23
14	0.27	0.23	30	0.25	0.06
15	0.22	0.14	31	0.18	0.12
16	0.28	0.08	32	0.58	0.06

Results

Item analysis

PCTc

The results of the item analysis for PCTc are presented in Table 1, in which the probability of valid solutions ranges from 0.02 to 0.88, and the item-total correlation ranges from −0.02 to 0.24, which are all below the acceptable criterion of 0.3.

PCTd

The results of the item analysis for PCTd are presented in Table 2. Firstly, for the fluency score, the mean values range from 5.31 to 7.31, and its item-total correlations range from 0.25 to 0.88. Notably, if item 14 is deleted, the minimum value of item-total correlation increases to 0.66. Secondly, regarding the originality score, the mean values range from 0.1 to 0.31, and the item-total correlations range from 0.1 to 0.39, with 13 out of 16 items below 0.3. Thirdly, in terms of the ratio of originality to fluency, the mean values range from 0.02 to 0.05, and the item-total correlations range from 0.01 to 0.38, with 13 out of 16 items below 0.3. Additionally, it is important to note that, for all the items in both the originality and the ratio, their values of skewness are all greater than 1 and the values of kurtosis are all greater than 3. These high values suggest a significant deviation from a normal distribution.

Reliability

The Cronbach's alpha values for PCTc and the three dimensions of PCTd are as follows: For PCTc, it is 0.43, indicating relatively low internal consistency reliability. For PCTd, the fluency score is 0.95, which is very good and indicates high internal consistency reliability. However, the Cronbach's alpha of the originality (0.59) and the ratio (0.53) are not very high, suggesting a relatively low internal consistency reliability.

Validity

The correlations between the PCT (PCTc and PCTd) and the RAT and AUT are presented in Table 3. First, the PCTc shows no correlation with both the RAT and the AUT, suggesting a poor convergent validity but an acceptable discriminant validity. Second, the fluency score of PCTd exhibits a positive correlation with both the fluency and flexibility scores of AUT, but no correlation with the originality score of AUT. The originality and ratio scores of PCTd also show positive correlation with the fluency score of AUT, but no correlations with the flexibility and originality scores of AUT. In addition, all three dimensions of the scores for PCTd exhibit no correlation with the RAT, indicating an acceptable discriminant validity.

Discussion

The main research question of this study was to examine whether the PCT is a valid instrument for assessing creativity. The PCT consists of PCTc and PCTd, which are supposed to measure CT and DT, respectively. Starting with PCTc, its item-total correlations were all lower than 0.3, which is considered unacceptable, indicating a lack of correlation between the items and the trait being measured, and low item discrimination. The reliability of PCTc was also low (Cronbach's alpha <0.6). Even though PCTc aims to assess CT, the PCTc showed no correlation with the standard CT task (RAT). Therefore, the results indicate that PCTc does not assess CT.

As for PCTd, it consists of three subscores: fluency, originality, and the ratio of originality to fluency. None of these scores showed any significant correlation with

Table 2. The results of the item analysis for PCTd.

	Fluency					Originality					Ratio (Originality/Fluency)				
Item	Mean	SD	Skewness	Kurtosis	Item-total correlation	Mean	SD	Skewness	Kurtosis	Item-total correlation	Mean	SD	Skewness	Kurtosis	Item-total correlation
1	5.95	1.58	−0.42	−0.15	0.71	0.14	0.52	4.48	23.09	0.31	0.03	0.11	4.38	23.97	0.33
2	5.93	1.62	−0.41	−0.08	0.71	0.18	0.53	3.14	9.71	0.22	0.03	0.08	3.99	12.58	0.17
3	6.40	1.62	−0.91	0.61	0.77	0.13	0.45	3.56	11.62	0.22	0.02	0.09	3.72	23.11	0.21
4	6.46	1.64	−1.06	0.95	0.68	0.23	0.66	2.98	9.01	0.22	0.03	0.08	3.38	8.46	0.09
5	6.18	1.69	−0.73	0.19	0.77	0.11	0.44	3.82	13.39	0.16	0.02	0.07	3.27	15.41	0.17
6	6.08	1.67	−0.63	0.07	0.83	0.28	0.67	2.48	5.51	0.13	0.04	0.10	2.83	8.44	0.07
7	5.61	1.76	−0.28	−0.46	0.79	0.24	0.63	2.65	6.17	0.22	0.04	0.11	3.49	7.63	0.13
8	5.31	1.79	−0.06	−0.52	0.76	0.10	0.43	4.62	22.27	0.23	0.02	0.06	3.81	17.92	0.19
9	5.65	1.79	−0.34	−0.42	0.81	0.19	0.54	3.23	10.89	0.1	0.04	0.11	2.94	10.59	0.01
10	6.06	1.72	−0.64	0.06	0.73	0.31	0.69	2.29	4.41	0.15	0.05	0.11	2.59	7.48	0.28
11	6.05	1.74	−0.62	−0.18	0.66	0.31	0.69	2.29	4.41	0.32	0.05	0.11	2.84	5.81	0.30
12	6.13	1.72	−0.63	−0.22	0.75	0.29	0.74	2.60	6.43	0.26	0.05	0.13	3.61	10.13	0.22
13	6.96	1.59	−1.37	1.02	0.8	0.14	0.50	3.71	13.37	0.39	0.02	0.07	3.95	18.34	0.38
14	7.13	1.36	−1.50	1.64	0.25	0.14	0.49	3.39	10.02	0.17	0.02	0.06	3.22	10.02	0.17
15	6.20	1.68	−0.67	−0.11	0.68	0.25	0.68	2.76	7.58	0.27	0.04	0.10	3.35	6.94	0.22
16	5.87	1.84	−0.43	−0.55	0.73	0.14	0.49	3.73	13.90	0.14	0.02	0.09	2.56	17.94	0.07

Table 3. Coefficients and significance levels for tests of correlation between the picture concept task (PCT, PCTc is for convergent thinking, PCTd is for divergent thinking), remote associate task (RAT), and alternate use task (AUT) (flu = fluency, fle = flexibility, ori = originality, ratio = ratio of originality to fluency).

	PCTc	PCTd–flu	PCTd–ori	PCTd–ratio	RAT	AUT–flu	AUT-fle	AUT–ori
PCTc	1							
PCTd-flu	−0.05	1						
PCTd-ori	0.04	0.39**	1					
PCTd-ratio	0.05	0.23**	0.92**	1				
RAT	0.05	−0.11	−0.03	−0.03	1			
AUT-flu	−0.06	0.44**	0.23**	0.18*	0.16	1		
AUT-fle	−0.13	0.35**	0.06	0.02	0.13	0.73**	1	
AUT-ori	0.07	0.12	0.01	−0.02	0.23**	0.25**	0.43**	1

**Correlation is significant at the 0.01 level (2–tailed).
*Correlation is significant at the 0.05 level (2–tailed).

RAT, which implies acceptable discriminant validity. Fluency showed high item-total correlations (>0.65) for all items except item 14, and its Cronbach's alpha coefficient was also very high (0.95), indicating excellent internal consistency. Furthermore, although the fluency of PCTd was not correlated with AUT's originality, it showed a positive correlation with AUTs fluency and flexibility. These results suggest that the fluency score of PCTd could be an effective measure of the fluency of DT. However, the originality score of PCTd showed low item-total correlations and reliability. Although it was positively correlated with the fluency score of AUT, it has no correlation with AUT originality, which is the most important aspect for validating PCTd's originality, as well as the flexibility score of AUT. This lack of correlation between the originality score of PCTd and AUT might be a consequence of the different methods to score originality in these two tasks. Moreover, the method for scoring originality in PCTd is highly influenced by sample size. The larger the sample size, the less likely a given idea is rated as original, due to the increased likelihood of repeated answers. These results suggest that the originality score of PCTd does not correspond to the originality of DT, and its scoring methods need improvement. Finally, the ratio score showed low item-total correlations, reliability, and poor validity, likely influenced by the originality score as well as the inherent technical limitations of ratio scores (Cronbach, 1941).

Therefore, in general, the present study suggests that the PCTc and PCTd should not be recommended as measurements for CT and DT, respectively. As mentioned, the PCT falls under the category of the Fluid Reasoning Index of the WCIS, designed for assessing logical and quantitative reasoning. While the PCTc and PCTd may appear to align with the definitions of CT and DT, it's important to note that the original design and construction of the PCT did not intend it for the measurement of creativity. However, the results suggest that PCTd could be considered a promising measurement of the fluency of DT. Furthermore, with improvements in the method of scoring for originality, PCTd may have potential to become a good measurement of DT, but further studies are needed to explore this possibility by changing the scoring method for originality. For example, inspired by a previous study used AUT (Ritter & Ferguson, 2017), an idea mentioned by 5% of the respondents scores 2 points, while an idea mentioned by 10% scores 1 point, and anything mentioned by more is scored 0 points. The use of this method will reduce the impact of sample size on originality scores, compared to the method used in the current study. At the same time, the originality dimension of the AUT, which is used to measure convergent validity, can also adopt the same scoring method. This would allow for a more effective analysis and interpretation of the scores from both tests. Finally, although the correlation between the PCT and typical DT and CT measures is not very strong, it may still relate to other creativity-related assessment tools, such as self-reports and product-based measures. Future research could further explore these connections.

Finally, we acknowledge two fundamental issues in this study that may be subject to debate: (1) Sample size – Due to experimental constraints, we were only able to collect relatively small and gender-imbalanced sample sizes. As a result, we were limited to using Classical Test Theory (CTT) and were unable to conduct an analysis using Item Response Theory (IRT). This limitation affects the generalizability of the test results, as they lack strong sample independence. (2) Scoring method differences in validity analysis – A key issue in comparing PCT-originality and AUT-originality scores stems from the differences in their scoring methods. In this study, we adopted the conventional scoring method used in prior PCT research (Kuypers et al., 2016; Mason et al., 2019, 2021; Uthaug et al., 2019), where scores are assigned based on the frequency of a given response in the sample. In contrast, AUT-originality was assessed using Likert ratings by independent raters. The PCT scoring method itself has inherent logical issues, as it is highly sensitive to sample

size. However, given our research objective – to critically examine the validity of PCT as a creativity measure – we decided to retain the established PCT scoring approach despite its limitations. While it is possible to score AUT-originality in a manner similar to PCT (e.g., assigning scores based on response frequency), this method does not effectively capture originality. Rare but nonsensical responses could still receive high scores, potentially misrepresenting the originality of creativity. In contrast, using human raters to assess originality and averaging their ratings provides a more valid measure of originality. Although employing Likert ratings for AUT-originality introduces challenges in direct comparison with PCT-originality, adopting a frequency-based scoring method for AUT-originality would not be a theoretically sound alternative for our research question. If PCTd is to be developed into a standardized creativity test, future studies should increase sample sizes to allow for IRT analysis and enhance the reliability of results. Additionally, for improved comparability and interpretability, as mentioned above, we recommend scoring originality in both PCT and AUT using a percentile-based approach (Ritter & Ferguson, 2017; e.g. responses appearing in 10% of the sample receive 1 point, responses appearing in 5% receive 2 points). This would facilitate more meaningful cross-test comparisons and interpretations. Besides, another limitation is that the validity results are affected by the varying reliability estimates.

However, our research question was to assess whether the PCT is a valid measure of creativity, and we believe that the present findings speak to this issue, suggesting that the PCT may not be an effective measure. Additionally, since the PCT is categorized under the Fluid Reasoning Index in the WISC, we could have included other fluid reasoning tests as external measures to gain a more comprehensive understanding of what the PCT is truly assessing. Given that the primary goal of this study was to evaluate whether the PCT is a valid creativity test, we did not pursue this approach. Future research could explore this aspect of the PCT further.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by an Advanced Grant of the European Research Council (ERC-2015-AdG-694722) and a Double-100 Talent Grant of the Province of Shandong, China, to BH.

ORCID

Bernhard Hommel http://orcid.org/0000-0003-4731-5125

Data availability statement

Data will be made available upon request from the lead author.

References

Acar, S., Runco, M. A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 39–49. https://doi.org/10.1037/aca0000256

Akbari Chermahini, S., Hickendorff, M., & Hommel, B. (2012). Development and validity of a Dutch version of the remote associates task: An item-response theory approach. *Thinking Skills and Creativity*, *7*(3), 177–186. https://doi.org/10.1016/j.tsc.2012.02.003

Baer, J., & Kaufman, J. C. (2019). Assessing creativity with the consensual assessment technique. In I. Lebuda & V. P. Glăveanu (Eds.), *The Palgrave handbook of social creativity research* (pp. 45–61). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-95498-1_3

Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, *27*, 90–93. https://doi.org/10.1016/j.cobeha.2018.09.017

Cristobal, E., Flavián, C., & Guinalíu, M. (2007). Perceived e-service quality (PeSQ): Measurement validation and effects on consumer satisfaction and web site loyalty. *Journal of Service Theory & Practice*, *17*(3), 317–340. https://doi.org/10.1108/09604520710744326

Cronbach, L. J. (1941). The reliability of ratio scores. *Educational and Psychological Measurement*, *1*(1), 269–277. https://doi.org/10.1177/001316444100100121

Cropley, A. (2020). Definitions of creativity. In Mark A, Runco, Steven R., Pritzker (Eds.), *Encyclopedia of Creativity* (3rd ed., pp. 315–322). Academic Press.

Forthmann, B., Jankowska, D. M., & Karwowski, M. (2021). How reliable and valid are frequency-based originality scores? Evidence from a sample of children and adolescents. *Thinking Skills and Creativity*, *41*, 100851. https://doi.org/10.1016/j.tsc.2021.100851

Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 94–112. https://doi.org/10.1037/aca0000196

Gabora, L. (2018). The neural basis and evolution of divergent and convergent thought. In R. E. Jung & O. Vartanian (Eds.), *The Cambridge handbook of the neuroscience of creativity* (pp. 58–70). Cambridge University Press. https://doi.org/10.1017/9781316556238.005

Girn, M., Mills, C., Roseman, L., Carhart-Harris, R. L., & Christoff, K. (2020). Updating the dynamic framework of thought: Creativity and psychedelics. *Neuroimage*, *213*,

116726. https://doi.org/10.1016/J.NEUROIMAGE.2020.116726

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill. https://psycnet.apa.org/record/1967-35015-000

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Kuypers, K. P. C., Riba, J., de la Fuente Revenga, M., Barker, S., Theunissen, E. L., & Ramaekers, J. G. (2016). Ayahuasca enhances creative divergent thinking while decreasing conventional convergent thinking. *Psychopharmacology*, *233*(18), 3395. https://doi.org/10.1007/S00213-016-4377-8

Mason, N. L., Kuypers, K. P. C., Reckweg, J. T., Müller, F., Tse, D. H. Y., da Rios, B., Toennes, S. W., Stiers, P., Feilding, A., & Ramaekers, J. G. (2021). Spontaneous and deliberate creative cognition during and after psilocybin exposure. *Translational Psychiatry*, *11*(1), 1–13. https://doi.org/10.1038/s41398-021-01335-5

Mason, N. L., Mischler, E., Uthaug, M. V., & Kuypers, K. P. C. (2019). Sub-acute effects of psilocybin on empathy, creative thinking, and subjective well-being. *Journal of Psychoactive Drugs*, *51*(2), 123–134. https://doi.org/10.1080/02791072.2019.1580804

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mednick, S. A. (1968). The remote associates test. *Journal of Creative Behavior*, *2*(3), 213–214. https://doi.org/10.1002/j.2162-6057.1968.tb00104.x

Prochazkova, L., Lippelt, D. P., Colzato, L. S., Kuchar, M., Sjoerds, Z., & Hommel, B. (2018). Exploring the effect of microdosing psychedelics on creativity in an open-label natural setting. *Psychopharmacology*, *235*(12), 3401–3413. https://doi.org/10.1007/S00213-018-5049-7

Ritter, S. M., & Ferguson, S. (2017). Happy creativity: Listening to happy music facilitates divergent thinking. *PLOS ONE*, *12*(9), e0182210. https://doi.org/10.1371/journal.pone.0182210

Runco, M. A., Acar, S., & Cayirdaga, N. (2017). A closer look at the creativity gap and why students are less creative at school than outside of school. *Thinking Skills and Creativity*, *24*, 242–249. https://doi.org/10.1016/j.tsc.2017.04.003

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Said-Metwaly, S., Noortgate, W. V. D., & Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity: Theories – Research - Applications*, *4*(2), 238–275. https://doi.org/10.1515/ctra-2017-0013

Sattler, M. J., & Dumont, R. (2008). WISC-IV subtests. In J. M. Sattler (Ed.), *Assessment of children: Cognitive foundations* (5th ed.). Jerome M. Sattler Publisher.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (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*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Stevenson, C., Baas, M., & van der Maas, H. (2021). A minimal theory of creative ability. *Journal of Intelligence*, *9*(1), 9. https://doi.org/10.3390/jintelligence9010009

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Uthaug, M. V., Lancelotta, R., van Oorsouw, K., Kuypers, K. P. C., Mason, N., Rak, J., Šuláková, A., Jurok, R., Maryška, M., Kuchař, M., Páleníček, T., Riba, J., & Ramaekers, J. G. (2019). A single inhalation of vapor from dried toad secretion containing 5-methoxy-N, N-dimethyltryptamine (5-MeO-MeO) in a naturalistic setting is related to sustained enhancement of satisfaction with life, mindfulness-related capacities, and a decrement of psychopathological symptoms. *Psychopharmacology*, *236*(9), 2653–2666. https://doi.org/10.1007/S00213-019-05236-W

Wechsler. (2003). *The Wechsler intelligence scale for children* (4th ed.). Pearson.

Wießner, I., Falchi, M., Maia, L. O., Daldegan-Bueno, D., Palhano-Fontes, F., Mason, N. L., Ramaekers, J. G., Gross, M. E., Schooler, J. W., Feilding, A., Ribeiro, S., Araujo, D. B., & Tófoli, L. F. (2022). LSD and creativity: Increased novelty and symbolic thinking, decreased utility and convergent thinking. *Journal of Psychopharmacology*, *36*(3), 348–359. https://doi.org/10.1177/02698811211069113

Zanon, C., Hutz, C. S., Yoo, H., & Kim, J. (2016). An application of item response theory to psychological test development. *Psicologia Reflexão e Crítica*, *29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x

Zhang, W., Sjoerds, Z., & Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *Neuroimage*, *210*(January), 116572. https://doi.org/10.1016/j.neuroimage.2020.116572