Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Are Creative People Better than Others at Recognizing Creative Work?

Steven E. Stemler\textsuperscript{a,\,*}, James C. Kaufman\textsuperscript{b,1}

\textsuperscript{a} Wesleyan University, USA
\textsuperscript{b} University of Connecticut, USA

ARTICLE INFO

Keywords:
Creativity
Assessment
Ratings
Expertise

ABSTRACT

It is often assumed that people with high ability in a domain will be excellent raters of quality within that same domain. This assumption is an underlying principle of using raters for creativity tasks, as in the Consensual Assessment Technique. While several prior studies have examined expert-novice differences in ratings, none have examined whether experts’ ability to identify the quality of a creative product is being driven more by their ability to identify high quality work, low quality work, or both. To address this question, a sample of 142 participants completed individual difference measures and rated the quality of several sets of creative captions. Unbeknownst to the participants, the captions had been identified \textit{a priori} by expert raters as being of particularly high or low quality. Hierarchical regression analyses revealed that after controlling for participants’ background and personality, those who scored significantly higher on any of three external measures of creativity also rated low-quality captions significantly lower than their peers; however, they did not rate the high-quality captions significantly higher. These findings support research in other domains suggesting that ratings of quality may be driven more by the lower end of the quality spectrum than the high end.

1. Introduction

\textit{“Not everybody is perfect and I don’t think we should be looking for perfect people,”} – Simon Cowell

\textit{ (“Paula Abdul dodges about review”, 2005)}

Personnel selection is often based on the assumption that the most important part of being an evaluator is being able to select the best candidate. When one thinks of reality competitions such as \textit{American Idol}, it is easy to gravitate to the most famous winners from Kelly Clarkson to Carrie Underwood. Yet it is perhaps even more impressive to note the many stars who made it to the finals, such as Grammy and Oscar winner Jennifer Hudson, Grammy nominee Adam Lambert, and American Music Award winner Clay Aiken. Distinguishing the absolute best from a strong array of finalists is difficult, but being able to quickly recognize the many clearly unqualified applicants and eliminate them from the pool is an essential skill. Indeed, perhaps the biggest star to emerge from the original show was judge Simon Cowell, whose sharp yet insightful critiques of contestants often went viral. What are the most important elements needed to successfully evaluate talent?

\textsuperscript{*} Corresponding author at: 207 High Street, Middletown, CT, 06437 USA.

\textit{E-mail addresses:} steven.stemler@wesleyan.edu (S.E. Stemler), james.kaufman@uconn.edu (J.C. Kaufman).

\textsuperscript{1} Neag School of Education, 2131 Hillside Road, Unit 3007, Storrs, CT 06269-3007, USA.

https://doi.org/10.1016/j.tsc.2020.100727

Received 10 July 2020; Received in revised form 31 August 2020; Accepted 3 September 2020

Available online 15 September 2020

1871-1871/© 2020 Elsevier Ltd. All rights reserved.
Thinking Skills and Creativity 38 (2020) 100727

S.E. Stemler and J.C. Kaufman

The ability to discern high versus low ability is a core requirement for a leader (Mitchell & Reiter-Palmon, 2017; Mumford, Hunter, Eubanks, Bedell, & Murphy, 2007). What does it take to be a successful evaluator? Within the creativity literature, there is a great deal of research on this topic. Researchers have studied whether people agree in their evaluation of creative work for over a century (Cattell, Glascoc, & Washburn, 1918), but Amabile (1982) brought this topic to the spotlight in her development of the Consensual Assessment Technique (CAT). In the CAT, a body of work (such as poems, drawings, or math equations) is evaluated by raters with an appropriate level of expertise within that domain. They assign scores to each product based on their own definitions of creativity. At no time do they consult with each other; discussion can lead to one rater’s opinions dominating (Hekkert & Van Wieringen, 1998). In addition, each piece is compared to the others as opposed to an ideal (Amabile, 1996).

What qualifies someone as a good rater of creative work? Pure novices do not show the same levels of reliability as experts, nor do their ratings always correspond with expert ratings (Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Baer, & Cole, 2009). By contrast, over many dozens of studies, expert raters have been found to agree at strikingly high rates (Baer, 1998; Baer, Kaufman, & Gentile, 2004; Baer, Kaufman, & Riggs, 2009). Further, this agreement is found even with expert raters across different cultures (Hennessey, Kim, Guomin, & Weili, 2008; Rostan, Pariser, & Gruber, 2002). The distance between novice and expert raters increases with domains that are less popular to the general public (i.e., the novice-expert gap was larger for poetry than for short stories; Kaufman et al., 2008, 2009). However, even a popular domain such as the cinema shows strong differences between different types of expert ratings (as measured by critical evaluation and peer-voted Academy Awards) and the more populist box office metrics (Simonton, 2009a, 2009b). In addition, experts (especially those who are trained in a domain, beyond simply demonstrating creative prowess) see creative work differently. Experts generally rate artwork as being more complex and interesting than novices (Locher, Smith, & Washburn, 1918), but Amabile (1982) brought this topic to the spotlight in her development of the Consensual Assessment Technique (CAT). In the CAT, a body of work (such as poems, drawings, or math equations) is evaluated by raters with an appropriate level of expertise within that domain. They assign scores to each product based on their own definitions of creativity. At no time do they consult with each other; discussion can lead to one rater’s opinions dominating (Hekkert & Van Wieringen, 1998). In addition, each piece is compared to the others as opposed to an ideal (Amabile, 1996).

Moving beyond the expert-novice dichotomy, solid agreement and reliability were found with a middle category of “quasi-experts” (Kaufman & Baer, 2012). Quasi-experts can include advanced students with experience in either teaching or creativity research (Kaufman, Baer, Copley, Reiter-Palmon, & Sinnett, 2013), novices who are highly gifted in the domain (Kaufman, Gentile, & Baer, 2005), and people with an intense interest in the topic (Plucker, Holden, & Neustadter, 2008; Plucker, Kaufman, Temple, & Qian, 2009).

Besides some form of expertise, what makes people better raters of creative works? Much of the research focuses on people evaluating their own ideas, which taps into the related construct of creative metacognition (Kaufman & Beghetto, 2013). Many studies have found that people who are better divergent thinkers are also better at evaluating their own ideas (Runco & Dow, 2004; Silvia, 2008). In addition, intelligence is also associated with accurately assessing one’s own ideas (Karwowski, Czerwonka, & Kaufman, 2018).

Evaluating one’s own ideas compared to evaluating other people’s ideas is a different ability, however, and studies that have looked at people evaluating both their own ideas and other people’s ideas have found differences. For example, people higher on divergent thinking were more accurate when rating their own ideas — but there was no difference when rating other people’s ideas (Runco & Smith, 1992). Similarly, people are more likely to overestimate their own creativity as opposed to the creativity of others (Grogeman, Wodniecka, & Klusak, 2006). In a reversed trend, people higher on openness to experience were more accurate when evaluating other people’s creativity, whereas there was no difference when they rated their own work (Birney, Beckmann, & Seah, 2016). Another study found that more creative people tended to generally underestimate other people’s creativity (Guo, Ge, & Pang, 2019). There have been few in-depth studies of creative evaluation skills and associated individual differences. Benedek et al. (2016), as part of two studies developing a creative evaluation test, found that higher ability in discerning creative work was associated with higher divergent thinking. In addition, they found women were slightly better at creative evaluation. The authors also found a negative evaluation bias — in other words, systematically underrating creative work. Both cognitive ability and openness were associated with less of a negative evaluation bias. Tan et al. (2015) studied how raters from three different majors (psychology, art, and computer science) evaluated Lego creatures built by children. They found no differences in how the three majors evaluated the works, although raters high in agreeableness were more lenient. These studies have different goals. Some of the studies with raters are aimed at refining the parameters of the CAT so creative products can be examined as outcomes. Other investigations are designed to take a metacognitive perspective on whether more creative people are also more accurate. Studies that specifically focus on evaluative abilities often use publicly acknowledged great artwork as a barometer of a rater’s skills. These are all important areas of scholarship; however, we suggest that an understudied component is examining not only how people differ or agree in assessing top work but also the ways in which they evaluate less creative outputs.

There are many reasons why people may vary in their selection of the most creative products. It is comparable to how easy it is to legitimately disagree about ability or quality among any small group of high performers, whether voting for Best Picture at the Academy Awards or selecting which job candidate to hire out of the three people who visited and delivered job talks. Any creative judgment involves subjectivity, but when distinguishing among a variety of works that already meet a high level, ultimate decisions of which participant or product is the best may be driven largely by raters’ individual interests and preferences. As a result, greater variability may be introduced into the rating process. In contrast, the first step to winnowing down a large pool of applicants or submissions is to be able to eliminate the obviously uncreative contributions. A good journal editor needs to have the ability (and confidence) to desk reject the multitudes of papers that are unoriginal or do not represent any type of advancement for the field. Consequently, we argue that it is equally (if not more) important for raters to consider focusing their attention on discerning the worst of the batch.

Past research has examined the broad question of raters’ ability to identify the best option vs. the worst in terms of quality. In a study of teachers’ practical intelligence, Elliott, Stemler, Grigorenko, Sternberg, and Hoffman (2011) found that teachers in the UK with more
than five years of teaching experience were better able to identify the worst response options to hypothetically challenging teaching scenarios than were teachers with less experience. Interestingly, however, the experienced teachers were no better than the novices at identifying the best responses (as determined by expert consensus). Building on this work, Stemler, Aggarwal, and Nithyanand (2016) created a series of situational judgment tests (SJTs) for salespeople and managers in India. They administered these tests in a series of three different studies and found that in every instance, those salespeople and managers who were rated more highly by their supervisors and who were more successful at meeting their sales targets were significantly better able to identify the worst response options on the SJTs than their peers, but no better or worse at identifying the best response options. Furthermore, the authors found that when traditional scoring approaches (in which the ability to identify the best and worst options were combined into a single score) were compared to scoring approaches that broke these skills apart, it was the ability to identify the worst response that was a significant predictor of performance and that drove the statistical significance of the traditional approaches to scoring in predicting key outcomes.

The aforementioned results suggest that it is not enough to simply look at general accuracy and interrater reliability overall as is typically done in research. Rather, it is worthwhile to investigate precisely where the agreement is coming from. In other words, is rater agreement uniform across the entire range of quality or is agreement is stronger at the lower or higher end of the spectrum? And will this pattern also be found in ratings of creativity?

The goal of the current study was to examine whether people who score highly on external measures of creativity are better able than others to identify work that has been classified a priori by expert consensus as being low or high in creative quality. Based on the prior literature, we expect that people who are highly creative themselves will be significantly better at spotting uncreative work, but will be no better than others at spotting work that it is judged as highly creative. The specific research hypotheses are listed below:

**H1.** People who score higher on any of three external measures of creativity (specifically, adaptability, mental flexibility, and writing creative captions) will be significantly better than others at identifying the least creative captions generated by other people on a creative captioning task.

**H2.** Scores on the three external measures of creativity will significantly predict the ability to detect the least creative responses, even after taking into account relevant demographic variables, and personality factors.

2. Methods

2.1. Sample

A sample of 142 undergraduate students enrolled in an introductory psychology course at small liberal arts university in the Northeastern US participated in the study. Participants received course credit for their participation and all the study procedures were approved by the Psychology Department’s Ethics Review Board.

Students had to be at least 18 years old in order to participate in the study, and the majority of the sample (n = 105, 75%) were in their first year of college. Across the full sample, 47 participants (33%) identified as male, 89 participants (63%) identified as female, 1 identified as transgender, 2 identified as not affiliated with a gender and 3 did not respond.

In terms of ethnic composition of the sample, 89 participants (63%) identified as White/European American, 18 (13%) identified as Asian/Asian American, 10 (8%) identified as Latinx/Hispanic-American, 7 (5%) identified as Black/African American, and 6 (4%) identified as biracial. The remaining 7% of participants either selected “other,” preferred not to answer, or skipped this item. Participants completed all tests below as part of a larger test battery.

A total of 123 (88%) of the participants reported that English was their primary language and 10 participants (7%) reported that they typically receive extended time accommodations on their testing.

2.2. Procedures

Participants were brought to a computer lab and tested in small group settings with a maximum of 25 people. Each participant used a desktop computer to complete the study. Each participant was randomly assigned an ID number as they checked in. The ID number was used to match participants’ paper-based results with their computer based questionnaires. Because the test battery involved a mixture of paper-pencil and online tests as well as a mixture of timed and untimed tests, the order of administration of the instruments was kept consistent for all participants. After first providing informed consent, the order of administration of the instruments was as follows: 1) Creative Picture Test, 2) Word Recognition Test, 3) Creative Caption Test and Ratings, 4) Personality Test, and 5) Demographics Questionnaire followed by a debriefing page.

Because one of the key outcomes of interest was participants’ ratings of the creativity of the captions of other people, we programmed the online survey to randomize the order in which those captions appeared for each participant.

The time it took participants to complete the online portion of the study ranged from 20 minutes to 49 minutes, with most people taking about half an hour to complete the computerized portion.

2.3. Measures

2.3.1. Tests of creativity

2.3.1.1. Creative picture test of adaptability. The Creative Picture Test of adaptability is the only paper-based instrument in this study.
It is similar to the Figural version of the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966; Torrance & Ball, 1998); however, Stemler has argued that this particular test is better conceived of as a measure of adaptability rather than divergent thinking (Stemler, 2009; Stemler et al., 2017). He argues that whereas divergent thinking involves the generation of new ideas, adaptability involves working around an existing constraint. Empirical data have shown low correlations between the Figural and Verbal tests of the TTCT (r = .06; Cramond, Matthews-Morgan, Bandalos, & Zuo, 2005), lending support to the notion that the figural and verbal tests are not measuring the same construct; however, whether one conceives of the test as a measure of figural divergent thinking or adaptability, either way it remains a measure of some aspect of creativity. Its criterion-related validity will be further examined in relation to two other more well-established measures of creative thinking included in the testing battery as well as a measure of personality.

Participants were asked to draw the most creative picture they could that integrated an existing rectangle and a curved line on the paper and then to briefly explain what they had drawn (see Fig. 1). Participants were given 6 minutes to complete the test.

In order to score the Creative Picture Test of adaptability, five undergraduate research assistants participated in a formal training session conducted by the lead author. The 45-minute training session involved reviewing the scoring rubric and completing evaluations on 10 sample items.

Raters generated a score for each test on four dimensions. The first dimension concerned each rater’s perception of the overall creativity of the image and explanation. Raters were asked to look at all 10 images in their training stack. Then, after they reviewed each image, they were asked to rate each of the images and their accompanying explanations for the overall level of creativity on a 1-9 scale. They were advised that a useful way to think about the task was to think about which of three boxes each image fell into: 1) high creativity, 2) moderate creativity, 3) low creativity. From there, they could consider whether each image fell in the high, moderate, or low level within each category. For example, one image might be thought to fall into the “high” creative box (which would consist of a score of 7, 8, or 9). The raters then had to determine whether within that “high” category the image was on the lower end (i.e., score of 7), moderate, (i.e., score of 8) or high end of the high category (i.e., score of 9).

After independently completing the ratings for each of the 10 training images, the research team then discussed their scores and thinking process via a modified version of the concurrent think-aloud protocol (Ericsson & Simon, 1980) in order to get a general sense of how people were thinking about the items as they scored them. After this process was complete, we moved onto the second phase in which raters were asked to follow the same procedure, but only attend to the Originality dimension (i.e., statistical rarity) of the image without regard to any other dimension. The raters then talked through any major discrepancies. This process was then repeated for the dimension of Task Appropriateness and for the dimension of Elaboration. No statistical evaluations of rater reliability were made during the training.

Finally, the raters were asked to engage in the same thinking process but to complete the ratings independently and at their own pace for all 142 images in the sample. The five research assistants worked independently and returned their ratings via email within 72 hours and these data were evaluated for interrater reliability.

The interrater reliability was assessed using a two-way mixed, absolute, average-measure intraclass correlation (McGraw & Wong, 1996) to assess the degree to which coders were providing the same values on their ratings of participant creativity in their captions. The resulting ICCs for each of the four dimensions were all in the excellent range (values greater than .74 = excellent; Cicchetti, 1994). The ICC values were as follows: Overall Creativity ratings (ICC = .86 consistency, .85 absolute), Originality ratings (ICC = .87 consistency, .86 absolute), Task Appropriateness ratings (ICC = .90 consistency, .88 absolute), and Elaboration ratings (ICC = .87 consistency, .86 absolute) indicating that coders had an extremely high degree of agreement and suggesting that creativity was rated similarly across coders. Consequently, it is appropriate to simply average the scores of the raters to arrive at a single score for each participant on each of these dimensions. Pearson correlations among the four rated dimensions revealed that Overall Creativity ratings were highly correlated with ratings of Originality (r = .84, p < .001), Task Appropriateness (r = .71, p < .001), and Elaboration (r = .86, p < .001). As a result, only the Overall Creativity rating was used for subsequent data analyses.

2.3.1.2. Word recognition test of mental flexibility. A previously published test of mental flexibility (Matthew & Stemler, 2013) was administered online to all participants. The test consists of four sentences in which the spelling of various words is mixed up, and the participants are asked to write out the words that they recognized in each sentence as quickly as they can (see Fig. 2). The word length ranged from 4 to 12 letters and the sentence length ratio (mixed up words/total words in sentence) ranged from 6/13 to 12/19.

The order of the sentences was randomized, and participants were given 4 minutes to complete the test. The test is scored according to a rubric in which longer words were assigned higher scores. For example, one point was assigned to correct words with 4-5 letters, and five points assigned to words with 12 letters. An aggregate score for the test was obtained by adding together the total score for correct rearranged words in each sentence. The test has been shown to have strong psychometric properties. In prior research (Matthew & Stemler, 2013), the internal consistency reliability estimate was alpha = .75 and the test has exhibited both significant predictive and incremental validity for predicting college grades over and above measures of fluid and crystallized intelligence (as measured by Ravens Advanced Progressive Matrices and the Mill Hill vocabulary test, respectively). Further, the test correlated significantly, albeit weakly, with the Torrance Test of Creative Thinking (r = .13, p < .05), suggesting that it was tapping a separate element of creativity. The internal consistency reliability of this 36-item test for the current sample of 142 participants was Cronbach’s alpha = .78.

2.3.1.3. Creative captioning test. The creative captioning test (e.g., Kaufman, Lee, Baer, & Lee, 2007) requires participants to write the most creative captions they can to accompany a visual prompt. Participants were given up to 5 minutes to write each caption, and they
Fig. 1. Creative Picture Test of Adaptability.
Prompt: Rhesareecrs pferer atsinstas who are scsnteiounois and rbelliae with sontrg iertesnt in the flied.

Answer: Researchers prefer assistants who are conscientious and reliable with strong interest in the field.

Fig. 2. Sample Word Recognition Test of Mental Flexibility Sentence.

were instructed to consider originality and relevance in order to be creative.

The participants’ captions were rated by six independent raters who were quasi-experts in the area of creativity (i.e., graduate students and advanced undergraduates studying and participating in research on the topic). The interrater reliability was assessed using a two-way mixed, absolute, average-measure intraclass correlation (McGraw & Wong, 1996) to assess the degree to which coders were providing the same values on their ratings of participant creativity in their captions. The resulting ICCs were generally in the range considered to be excellent (i.e., .74 = excellent, Cichetti, 1994) for all three prompts. The first image prompt in the study involved a person sitting at the bottom of a stairway (ICC = .83 consistency, .82 absolute), the second image prompt involved a man sitting at a desk (ICC = .78 consistency, .72 absolute), and the third image prompt showed a group of students eating french fries (ICC = .82 consistency, .78 absolute), indicating that coders had a very high degree of agreement indicating that creativity was rated similarly across coders (Hallgren, 2012).

2.3.2. Personality

2.3.2.1. The aspiring minds personality inventory. The Aspiring Minds Personality Inventory is a self-report personality instrument designed to yield scores for each of the personality traits associated with the five factor model of personality (McCrae & John, 1992). Participants responded using a 5-point Likert scale describing how accurately each statement applied to them (1 = very inaccurate, 5 = very accurate). The results of the AMPI scales have been shown to have concurrent validity with more well-established measures of the Big Five, such as the 50-item IPIP version of the Big Five Markers (Goldberg, 1992). A recent study by Stemler and Aggarwal (under review) showed the correlations between the aforementioned IPIP measures of the Big Five and the AMPI in a sample of 132 college students in the United States to be as follows: Openness (.36, p < .01), Conscientiousness (.55, p < .01), Extraversion (.62, p < .01), Agreeableness (.58, p < .01), and Neuroticism (.68, p < .01).

The AMPI scales had acceptable internal consistency reliability estimates in the current sample: Openness (n items = 11, alpha = .74); Extraversion (n items = 11, alpha = .62); Agreeableness (n items = 12, alpha = .74); and Neuroticism (n items = 12, alpha = .78), with the exception of the Conscientiousness subscale (n items = 11, alpha = .41), which had far lower reliability for this sample than for other samples using this same instrument (c.f., the concurrent validity sample of 132 participants exhibited an alpha = .60 for Conscientiousness). A principal components analysis reveals that the five factors explain 36% of variance in the data.

2.3.3. Dependent variables

2.3.3.1. Best, Worst, and Moderate captioning scales. After writing their caption, participants were then presented with 10 real captions that had been provided by participants in a previous study. The ten captions that were chosen for inclusion in the study were taken from a previous study and were deliberately intended to represent captions that had been rated by experts as corresponding to high, medium, and low creativity.

The first image prompt (person sitting at the bottom of stairs) was captioned by 1,520 participants whose responses were rated by 14 expert raters and had an average absolute intraclass correlation coefficient of reliability = .91 (consistency ICC = .93). The ratings, which ranged from 1 (not at all creative) to 6 (highly creative) for each caption were then averaged across all 14 raters. The average ratings were plotted and the distribution was divided into thirds based on the results. Average scores from 1-1.78 constituted the bottom third of the ratings, scores from 1.79-2.4 constituted the middle third, and ratings higher than 2.4 constituted the top third of ratings. We choose three captions with average ratings of 2.5 or greater to represent the “Best”, three with average ratings of 1.78 or lower to represent the “Worst” and four with average ratings between 1.79-2.4 to represent “moderate” creative captions.

The same process was repeated for the other two image prompts. For Caption #2 (Man at Desk), eight previous expert raters had an absolute ICC of .80 (consistency ICC = .82) for 577 captions. For Caption #3 (Friends eating french fries), a total of 5,599 responses were rated by four expert raters previously who had an average absolute intraclass correlation coefficient = .73 (consistency ICC = .74). All of these ICC values were in the excellent range (Cichette, 1994).
For each of these captions, participants in the present study were asked to rate the quality of the 10 captions per image on the same 1 (not very creative) to 6 (highly creative) scale. In order to create a new scale, the ratings given by participants in the current study on just those items that were identified as being the “Best” in the previous study (i.e., the top 3 most highly rated items for each prompt by expert raters from the previous study) were averaged\(^2\) to yield a “Best” score for each participant (9 items, scale alpha = .79). In a similar manner, the participant ratings of the nine caption items that were chosen as the “Worst” across the three prompts (the bottom 3 least highly rated items for each prompt by external raters) were averaged to yield a “Worst” score for each participant in the current study (9 items, scale alpha = .65). In addition, participants’ ratings of the 12 captions that were chosen as “Moderate” across the three prompts were averaged to yield a “Moderate” score for each participant in the current study (12 items, scale alpha = .67).

These three scales then served as the outcome variables in the subsequent regression analyses. The following factors were regressed on these outcomes: Background variables of participants (gender, English as primary language, test accommodations); Big-5 Personality traits, along with the following three measures of creativity: 1) Adaptability (Creative Picture Test), 2) Mental Flexibility (Word Recognition Test), and 3) External ratings of the participants’ own caption to the exact same task.

### 3.2. Data Analyses

Sample sizes vary slightly due to skipped items on some measures. Hypothesis testing was conducted using correlation analysis and hierarchical regressions. Descriptive statistics for all measures are presented in Table 2.

Table 3 presents the Pearson correlation coefficients for all variables under investigation. Note that the relevant reliability values are reported in the diagonals.

The results of the correlational analyses presented in Table 3 support Hypothesis #1 that participants who scored higher on any of the three measures of creativity tend to be significantly better than others at identifying uncreative work. Specifically, participants who scored higher on the Creative Picture Test of adaptability, the Word Recognition Test of mental flexibility, and participants whose own creative captions were scored higher by external raters gave significantly lower ratings than their peers to captions that were rated \textit{a priori} by experts as not creative, but there were no significant differences from others in their ratings of captions that were identified \textit{a priori} by the experts as highly creative.

The data further revealed that participants scoring higher on the personality characteristics of openness and agreeableness gave significantly higher ratings than their peers to captions that had been rated \textit{a priori} by experts as being highly creative, but participants

\footnote{The mean of the 9 ratings were used for each participant rather than simply summing the ratings in order to account for any missing values that resulted from a participant skipping a particular rating. The data analyses were run both ways (using sums v. using means) and the results did not differ substantively.}
who scored higher on these personality dimensions did not differ significantly from others with regard to their ratings of captions identified a priori as not creative. Female participants showed a similar pattern as well, with females giving significantly higher ratings to work that had been identified a priori by quasi-experts as being of high quality.

Finally, the correlation results reveal that participants scoring higher on the Creative Picture Test of adaptability gave significantly lower ratings to items on the scale (\(r = -0.27, p < .01\)) than their peers to captions that were identified a priori by experts as uncreative. Note that negative values correspond to lower ratings given by participants while positive values suggest higher ratings given to captions that are classified by experts as uncreative.

The results of this regression show that after controlling for demographic variables and personality factors, participants’ own level of creativity was a statistically significant predictor of their ratings of captions that were a priori identified by experts as uncreative, accounting for about one-fifth of the total variance in ratings. Specifically, all other things being equal, individuals who scored higher on the Creative Picture Test of adaptability gave significantly lower ratings (Beta = -0.27, p < .01) than their peers to captions that were identified by experts as being uncreative.

Table 4 presents the results of the regression analysis by looking at which variables predict participants’ ratings of captions that were identified a priori by experts as being uncreative. Table 5 presents the results of the regression analysis by looking at which variables predict scores on the “Best” scale, which takes the mean of participants’ ratings of items that were identified by experts a priori as being highly creative.

---

Table 1
Regression Results for Ratings of Overall Creativity Across all 30 items.

| Variables                  | R   | R² | Change Statistics |
|----------------------------|-----|----|-------------------|
|                            |     |    | F                 |
|                            |     |    | Change          |
|                            |     |    | Sig              |
|                            |     |    | 0.15             |
|                            |     |    | 0.83             |
|                            |     |    | .06, ns          |

Notes: * p < .05, ** p < .01, *** p < .001

Table 2
Descriptive Statistics for All Scales.

| Measures                          | N   | Minimum | Maximum | Mean | SD  |
|-----------------------------------|-----|---------|---------|------|-----|
| 1. Worst Scale                    | 141 | 1.0     | 3.2     | 1.7  | .48 |
| 2. Moderate Scale                 | 141 | 1.3     | 4.6     | 2.9  | .60 |
| 3. Best Scale                     | 141 | 1.4     | 6.0     | 4.1  | .85 |
| 4. Creative Captions              | 141 | 4.2     | 13.7    | 8.6  | 2.10|
| 5. Adaptability                   | 135 | 1.0     | 8.8     | 5.1  | 1.60|
| 6. Mental Flexibility             | 141 | 15.0    | 36.0    | 27.1 | 4.19|
| 7. Openness                       | 140 | 18.0    | 55.0    | 43.9 | 5.90|
| 8. Conscientiousness              | 140 | 19.0    | 42.0    | 32.5 | 5.41|
| 9. Extraversion                   | 140 | 12.0    | 53.0    | 40.8 | 5.97|
| 10. Agreeableness                 | 140 | 21.0    | 60.0    | 48.4 | 6.15|
| 11. Neuroticism                   | 140 | 20.0    | 58.0    | 41.1 | 7.78|
| 12. Gender                        | 135 | 0.0     | 1.0     | 0.7  | 0.48|
| 13. English as Primary Language   | 140 | 0.0     | 1.0     | 0.9  | 0.33|
| 14. Extended Time                 | 140 | 0.0     | 1.0     | 0.1  | 0.26|

Note: Gender 0=Male, 1=Female; English as Primary Language 0=No, 1=Yes; Extended Time 0=No, 1=Yes
### Table 3
Full Correlations for All Measures and Background Variables.

| Scale | 1) | 2) | 3) | 4) | 5) | 6) | 7) | 8) | 9) | 10) | 11) | 12) | 13) |
|-------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1. Worst Scale | 0.65 | | | | | | | | | | | | |
| 2. Moderate Scale | 0.40 | * | 0.67 | | | | | | | | | | |
| 3. Best Scale | 0.15 | | 0.65 | * | 0.79 | | | | | | | | |
| 4. Creative Captions | -0.28 | * | -0.14 | -0.08 | 0.78 | | | | | | | | |
| 5. Adaptability | -0.35 | * | -0.22 | * | 0.06 | 0.18 | * | 0.85 | | | | | |
| 6. Mental Flexibility | -0.28 | * | -0.05 | 0.02 | 0.30 | ** | 0.18 | * | 0.78 | | | | |
| 7. Openness | -0.01 | 0.06 | 0.27 | ** | 0.08 | 0.03 | * | 0.09 | 0.74 | | | | |
| 8. Conscientiousness | 0.02 | -0.02 | 0.14 | -0.06 | 0.05 | 0.14 | 0.03 | 0.41 | | | | | |
| 9. Extraversion | -0.04 | 0.05 | 0.14 | 0.11 | 0.08 | -0.02 | 0.14 | 0.19 | * | 0.62 | | | |
| 10. Agreeableness | 0.09 | 0.11 | 0.32 | ** | 0.01 | 0.03 | 0.05 | 0.59 | * | 0.15 | 0.35 | * | 0.74 | |
| 11. Neuroticism | 0.00 | * | -0.05 | -0.03 | -0.04 | -0.10 | -0.15 | 0.16 | -0.37 | ** | -0.16 | -0.03 | 0.78 | |
| 12. Gender | -0.05 | 0.08 | 0.27 | ** | -0.13 | 0.00 | -0.12 | 0.20 | * | -0.01 | -0.04 | 0.26 | ** | 0.22 | * | 1.00 | |
| 13. English Primary Lang | -0.01 | -0.03 | -0.01 | 0.13 | 0.02 | 0.15 | -0.04 | -0.07 | 0.23 | ** | 0.04 | -0.02 | -0.12 | 1.00 | |
| 14. Extended Testing | 0.20 | * | 0.18 | * | 0.07 | -0.10 | -0.18 | * | -0.12 | 0.03 | -0.07 | -0.13 | 0.05 | 0.08 | 0.02 | -0.07 | |

**Note:** * p < .05; ** p < .01; N = 141; Gender 1 = Female; English as Primary Language 1 = Yes; Extended Time 1 = Yes
Thinking Skills and Creativity 38 (2020) 100727

Table 4
Regression Results for the Ability to Spot the “Worst” Creative Captions.

| Dependent Variable: "Worst Captions" | R  | R2  | Change Statistics |
|-------------------------------------|----|-----|------------------|
|                                      |    |     | R2change | F  |
| 1                                   | 0.20 | 0.04 | 0.04   | 1.88 | ns  |
| 2                                   | 0.25 | 0.06 | 0.02   | 1.00 | ns  |
| 3                                   | 0.50 | 0.25 | 0.19   | 3.50 | *** |

Notes: * p < .05, ** p < .01, *** p < .001

The results of this regression show that all variables taken together explained a significant amount of variance in the data (R2 = 18%); however, there was only one statistically significant predictor variable. Specifically, females gave significantly higher ratings (Beta = .22, p < .05) than did males to captions that were identified a priori by experts as being highly creative. Importantly, there were no statistically significant difference between people who scored higher on measures of creativity and those who scored lower with regard to their ratings of the captions that were identified by experts as being highly creative.

Finally, Table 6 presents the results of the regression analysis by looking at which variables predict scores on the “Moderate” scale, which takes the mean of participants’ ratings of items that were rated by experts as not falling into either the top third or bottom third in terms of their creativity. The results reveal that the regression equation does not predict a statistically significant amount of variance in the data overall and therefore it would be inappropriate to interpret the individual regression coefficients.

Agreeableness and Openness both showed statistically significant correlations with the ability to identify the “Best” captions (r = .32 and r = .27 respectively). Yet neither of these variables were statistically significant predictors in any of the final regression equations. To examine this question in more detail, we conducted a median split for Agreeableness scores and Openness scores. We then compared people high in Agreeableness to those low in Agreeableness to determine whether they differed significantly in their ratings of the Best, Worst, and Moderately creative captions. We then followed the same procedure for Openness.

The results revealed that people who scored high on Agreeableness (above the median) gave significantly higher ratings to the “Best” responses (X = 4.33 SD = .85) than did those who scored low on Agreeableness (X = 3.92, SD = .81; t = 2.91, df = 138, p < .01, d = .49). However, people who were high on Agreeableness did not differ significantly in the ratings of the Moderate responses (t = 1.09, df = 138, p = .28) or Worst responses (t = 1.72, df = 138, p = .09), ruling out the possibility that highly Agreeable people simply gave higher responses to all captions across the board. Rather, there does seem to be some kind of relationship between Agreeableness and the ability to correctly identify the Best responses.

The ratings from people high on Openness, by contrast, were not significantly different from those scoring low on Openness with regard to their ability to identify the Best responses (t = 1.76, df = 138, p < .08), Moderate responses (t = 0.28, df = 138, p = .78), or Worst responses (t = 0.86, df = 138, p = .39).

Finally, we ran a 2 × 2 chi-square test of association between Gender (Male/Female) and Agreeableness score (High/Low). The results indicated a statistically significant association (X²1 = 6.75, p < .01) with greater numbers of females represented in high
Agreeableness condition in this sample, a finding that is consistent with prior literature in this area (Weisberg, DeYoung, & Hirsh, 2011).

4. Discussion

The results of this investigation show that participants who score higher on external measures of creativity give significantly lower ratings than their peers to work that was identified a priori by experts as uncreative. That is, they were more likely to agree with the experts that the captions were of low quality. At the same time, individuals scoring higher on external measure of creativity did not differ in any significant way in terms of their ratings of captions that were identified a priori by experts as being highly creative. Stated differently, whether or not you yourself had exhibited creative ability had no predictable association with how you rated work that had been deemed highly creative by experts in the field – but it mattered a great deal for how you rated work considered uncreative.

One potential interpretation of this finding would be that the individuals scoring higher on creativity perhaps use a broader range of scoring values on the rating scale whereas those at the lower end may have made less extreme evaluations, particularly at the lower end of the rating scale. However, if that were the case, the regression results would have shown a statistically significant relationship between overall ratings and scores on any of the three creativity measures, but they did not. To further test this hypothesis, we did a median split on each of the three creativity variables to create “high” v. “low” creativity groups. We then compared the overall ratings for each group using a Levene’s test for homogeneity of variance and found in each case that the variance was not significantly different. Thus, it is not the case that people with more creative ability are simply using a broader range of values on the rating scale than their peers.

Instead, it seems to be the case that individuals who score higher on objective tests of creativity are significantly more accurate (in terms of agreeing with experts) in their ratings of uncreative work than are their peers, but they are no more accurate than their peers in terms of identifying highly creative work. The findings of this study are in line with past research has examined the broad question of

| Dependent Variable “Best Captions” | R    | R2   | Change Statistics |
|-----------------------------------|------|------|--------------------|
|                                    | R2change | F   |
| 1                                 | 0.28 | 0.08 | 0.08 3.90 **       |
| 2                                 | 0.42 | 0.17 | 0.09 3.27 **       |
| 3                                 | 0.42 | 0.18 | 0.01 2.34 **       |

| Dependent Variable: “Best Captions” | B    | SE   | Beta | t    | Sig |
|-------------------------------------|------|------|------|------|-----|
| STEP 1:                             |      |      |      |      |     |
| Gender                              | 0.49 | 0.15 | 0.28 | 3.26 | 0.00 *** |
| English as Primary Language         | 0.08 | 0.22 | 0.03 | 0.38 | 0.71 |
| Extended Time                       | 0.22 | 0.28 | 0.07 | 0.80 | 0.42 |
| STEP 2:                             |      |      |      |      |     |
| Gender                              | 0.40 | 0.16 | 0.23 | 2.58 | 0.01 * |
| English as Primary Language         | 0.04 | 0.22 | 0.02 | 0.20 | 0.84 |
| Extended Time                       | 0.25 | 0.27 | 0.08 | 0.92 | 0.36 |
| Openness                            | 0.02 | 0.02 | 0.13 | 1.30 | 0.20 |
| Conscientiousness                   | 0.02 | 0.01 | 0.09 | 1.05 | 0.30 |
| Extraversion                        | 0.01 | 0.01 | 0.06 | 0.69 | 0.49 |
| Agreeableness                       | 0.02 | 0.02 | 0.14 | 1.27 | 0.21 |
| Neuroticism                         | −0.01| 0.01 | −0.05| −0.59| 0.56 |
| STEP 3:                             |      |      |      |      |     |
| Gender                              | 0.39 | 0.16 | 0.22 | 2.38 | 0.02 * |
| English as Primary Language         | 0.05 | 0.23 | 0.02 | 0.22 | 0.83 |
| Extended Time                       | 0.27 | 0.29 | 0.08 | 0.95 | 0.34 |
| Openness                            | 0.02 | 0.02 | 0.14 | 1.29 | 0.20 |
| Conscientiousness                   | 0.01 | 0.02 | 0.08 | 0.89 | 0.37 |
| Extraversion                        | 0.01 | 0.01 | 0.07 | 0.74 | 0.46 |
| Agreeableness                       | 0.02 | 0.02 | 0.14 | 1.18 | 0.24 |
| Neuroticism                         | −0.01| 0.01 | −0.05| −0.52| 0.61 |
| Word Recognition Test (Mental Flex) | 0.01 | 0.02 | 0.02 | 0.25 | 0.80 |
| Creative Picture Test (Adaptability) | 0.03 | 0.05 | 0.06 | 0.73 | 0.47 |
| Creative Captions                   | −0.03| 0.04 | −0.08| −0.87| 0.39 |

Notes: * p < .05, ** p < .01, *** p <.001

Agreeableness condition in this sample, a finding that is consistent with prior literature in this area (Weisberg, DeYoung, & Hirsh, 2011).
raters’ ability to identify the best option vs. the worst in terms of quality in the domains of teaching (Elliott et al., 2011) and sales/management (Stemler et al., 2016) which has consistently shown that the ability to identify the worst options was more predictive of successful outcomes than the ability to identify the best options. This discrepancy suggests that the research findings discussed in the introduction may benefit from additional studies and theoretical unpacking. Simonton (2018) provides an excellent start to such a theoretical unpacking by specifying seven specific ways that ideas can be thought of as uncreative. Future investigations may wish to examine how different types of raters agree or disagree when evaluating products that are notably higher or lower in creativity. Such studies may also want to examine what variables differentially predict rater agreement on high and low creative works. There are some studies that are related (albeit indirectly), although most emphasize how highly creative ideas are judged.

Some work has already examined how judges who differ in expertise or creativity evaluate products that are considered to reflect higher or lower creativity. However, it is the way that more original or creative work is rated which is typically the focus. Közbelt and Serafin (2009), for example, had experts and novices evaluate artwork over a period of time as the piece was completed. When experts evaluated the art, their ratings differed for high versus low creative work; highly creative drawings evolved in an unpredictable fashion, whereas less creative work developed more incrementally. When novices evaluated the drawings, no such differences were found. Caroff and Besançon (2008) had raters take a divergent thinking task and then rate advertisements. Raters who received higher originality scores assigned higher creativity ratings to advertisements considered more original.

Other studies have not examined different types of judges but have examined differences in how different types of creative work is evaluated. Ratings of task appropriateness, for example, were useful in helping predict separate creativity ratings when the ideas being evaluated were highly original (Diedrich, Benedek, Jauk, & Neubauer, 2015). Similarly, Licuanan, Dailey, and Mumford (2007) found that a negative evaluation bias was more likely to be found in very original ideas. In a different study, raters showed more disagreement when assessing tasks in which the creators were instructed to be creative, as opposed to tasks with neutral instructions (Forthmann et al., 2017). Given past studies on the significant impact of instruction (e.g., O’Hara & Sternberg, 2001), it is reasonable to assume that the tasks with “be creative” instructions were more likely to yield more creative responses.

### Table 6

Regression Results for the Ability to Spot the “Moderate” Creative Captions.

| Dependent Variable | R | R2 | R2Change | F | Sig |
|--------------------|---|----|----------|---|-----|
| 1                  | 0.19 | 0.04 | 0.04 | 1.70 | ns |
| 2                  | 0.24 | 0.06 | 0.02 | 0.92 | ns |
| 3                  | 0.33 | 0.11 | 0.05 | 1.30 | ns |

**Dependent Variable: “Moderately Creative Captions”**

| B     | SE  | Beta | t   | Sig |
|-------|-----|------|-----|-----|
| STEP 1: |
| Gender | 0.10 | 0.11 | 0.08 | 0.88 | 0.38 |
| English as Primary Language | −0.02 | 0.16 | −0.01 | −0.15 | 0.89 |
| Extended Time | 0.41 | 0.20 | 0.17 | 2.03 | 0.05 |

**STEP 2:**

| Gender | 0.10 | 0.12 | 0.08 | 0.83 | 0.41 |
| English as Primary Language | −0.06 | 0.17 | −0.03 | −0.37 | 0.71 |
| Extended Time | 0.42 | 0.21 | 0.18 | 2.04 | 0.04 |

| Openness | 0.00 | 0.01 | 0.00 | 0.02 | 0.99 |
| Conscientiousness | −0.01 | 0.01 | −0.06 | −0.67 | 0.50 |
| Extraversion | 0.01 | 0.01 | 0.05 | 0.53 | 0.60 |
| Agreeableness | 0.01 | 0.01 | 0.07 | 0.58 | 0.56 |
| Neuroticism | −0.01 | 0.01 | −0.10 | −0.97 | 0.34 |

**STEP 3:**

| Gender | 0.09 | 0.12 | 0.07 | 0.78 | 0.44 |
| English as Primary Language | −0.06 | 0.17 | −0.03 | −0.33 | 0.75 |
| Extended Time | 0.33 | 0.21 | 0.14 | 1.58 | 0.12 |
| Openness | 0.00 | 0.01 | 0.02 | 0.19 | 0.85 |
| Conscientiousness | −0.01 | 0.01 | −0.08 | −0.82 | 0.41 |
| Extraversion | 0.01 | 0.01 | 0.08 | 0.75 | 0.46 |
| Agreeableness | 0.01 | 0.01 | 0.06 | 0.49 | 0.62 |
| Neuroticism | −0.01 | 0.01 | −0.11 | −1.14 | 0.26 |
| Word Recognition Test (Mental Flex) | 0.01 | 0.01 | 0.04 | 0.44 | 0.66 |
| Creative Picture Test (Adaptability) | −0.08 | 0.03 | −0.20 | −2.22 | 0.03 |

**Notes:** * p < .05, ** p < .01, *** p < .001
With regard to the findings related to personality, although Agreeableness was not a statistically significant predictor of any of the outcome variables in the final regression models, there was evidence of a relationship with accuracy of ratings with regard to work considered by experts to be the most creative work. It is possible that this is an artifact of there being a greater proportion of women in the sample given that agreeableness is arguably the least associated with creativity out of the big five personality factors; if anything some studies have shown a negative relationship with agreeableness and self-reported creativity (Peist, 1998; Silvia, Kaufman, Reiter-Palmon, & Wigert, 2011). Nevertheless, the results from this study merit further investigation, potentially utilizing a more facet-based approach to personality assessment.

Why might females be more accurate at identifying the best creative works? One reason might be gender differences in analytical reasoning strategies. Females are more likely to solve a problem using an in-depth, algorithmic style, whereas males are more likely to use strategies that are quicker and unconventional (Gallagher & DeLisi, 1994). This discrepancy is one possible reason why females tend to perform better in mathematics classes than males, but the trend is reversed on timed standardized tests (Gallagher & Kaufman, 2005). Even though males’ preferred strategy has been considered more inventive (e.g., Fennema, Carpenter, Jacobs, Franke, & Levi, 1998), it may take more focus and effort for non-experts to more effectively rate creative work. Similarly, females have consistently been shown to perform slightly better on verbal creativity tasks, whereas males tend to show higher creativity in visuospatial creativity tasks (Baer & Kaufman, 2008). Relatedly, females are more likely to use verbal processing when performing a task (even a visuospatial one), whereas males tend to use visuospatial processing (Hugdahl, Thomsen, & Erslund, 2006). Given the product rated in this study was written, females may have had an inherent advantage. That said, such arguments would predict that females would be more accurate on all ratings. The reasons why gender only played a role in discerning the best creative products require further study.

Clearly, between the investigations discussed in the introduction and those highlighted above, there is a research foundation upon which to build. For example, although prior research has examined novice-expert differences (e.g., Kaufman & Baer, 2012) in the evaluation of creative work, the results of the present study suggest that we may want to focus more on what distinguishes ratings of less creative work where there is likely to be more agreement between these two groups. This direction stands in contrast to most past scholarship, which is more likely to focus on the higher end of the spectrum.

The use of sentence captions has both positive and negative implications. On one hand, there are no award competitions or employment opportunities for writing captions. The generalizability of evaluating sentence captions to evaluating potential job applicants is a reasonable question to be raised. On the other hand, given the extensive domain differences in how experts and novices agree (i.e., Kaufman et al., 2013), any specific domain would run the risk of being inapplicable to other areas. Sentence captions were chosen because they are as close to general domain that does not require much specific content knowledge (e.g., Kaufman et al., 2007).

4.1. Limitations of the Study

Certainly, the present study is only an initial investigation and has significant limitations. The participants who rated the captions were distinguished based on their scores on measures of creativity, personality, and background variables. Although this approach covers a wide swath, it does not use the expert/quasi-expert/novice paradigm that so many earlier studies have used. Next, the creativity being rated in this study pertained to image captioning. Future investigations may consider using a clearer domain (such as creative writing, art, or a STEM field). Third, the participants were asked to provide ratings on a 1-6 scale for each caption rather than choosing which captions they felt were most and least creative. Although this approach presents a useful starting point, future studies may want to consider forcing participants to rank order the captions in order to better evaluate the sensitivity and specificity of the ratings. Fourth, there is the issue of measurements chosen; although these instruments have shown strong relationships with better-established tests, they are still less commonly used (and the reliability of the conscientious personality factor was notably low, calling into question any analysis including that variable). Finally, it is important to note that there are other constructs (such as intelligence, critical thinking, or knowledge of the domain in question) that should also be examined in relation to differences in the ability to rate high, low, and moderate creativity. Future studies should also include additional well-established measures of these constructs, such as the Raven’s matrices or the Torrance Tests of Creative Thinking.

5. Conclusions

What do the results of this study mean in practical terms? In many real-life evaluation circumstances, the first step is to screen out the lesser entrees. For example, if there are 300 job candidates for a single position which requires creativity, the initial phase typically involves winnowing the list down to less than 15-20 candidates to actually examine in detail. The next stages (phone interviews, committee discussions, and eventually campus visits) are all part of the process of finding the best candidate. It is at this point that the selection strategy often flips from ruling out the least qualified to identifying the most qualified candidate. This change in strategy from initial screening to final selection may actually result in lower levels of rater reliability than could be achieved by continuing to employ the strategy of continuing to rule out the worst candidates or products until one arrives at a finalist.

Indeed, based on the results of the current study, one could reasonably argue that if you want to take full advantage of having people with a demonstrated ability in the area you are looking for on the selection committee, the place where that ability is most important is in getting rid of the weakest candidates, not in selecting the best. Considering the importance of creativity in many hiring opportunities and award selections, this change could have a significant impact. Everyone voting to eliminate the worst is a different dynamic than voting on the best. Given that interrater reliability results across numerous studies have now shown that agreement is consistently stronger when identifying low quality/worst options than identifying high quality/best options, it is likely that the most important predictor of a committee’s effectiveness may be in not selecting the bottom candidates to be finalists for a job/award as
opposed to trying to identify the best top finalists.

One prominent domain that relies on this reverse selection strategy is the legal system. When selecting members of a jury, attorneys on both sides have a say in which potential jury members they would like to rule out. However, they do not get to identify potential jury members that they definitely want to keep on the case. The assumption of this approach is that most of the people in the pool can do the job and do it well. As a result, the system simply needs to get rid of any jurors who might be of low quality (e.g., possess biases). The goal is not to find the ideal/wisest and most judicious individuals. The goal is instead to work backward and rule out the least qualified to serve as impartial members of a jury. The present study indicates that a comparable system may also work for selecting creative people or products.

Finally, if we look at the major issues our time, which encompass everything from choosing our political leaders to fighting a global pandemic to dealing with climate change and systemic racism, it is important for people to be able to identify ideal candidates, strategies, and solutions to problems. Yet, with the proliferation of all kinds of information online and the growing influence of social media, we are in a time when the people who are making decisions about such matters may not themselves possess sufficient levels of creativity to be able to differentiate the worst ideas from the best. When it comes down to the final rounds of decision-making where the choice is between picking amongst only a few options, a lack of creativity may not be a disadvantage, but that is only true if highly creative experts in a domain are involved at the early stages of decisions-making so that particularly poor solutions for solving our top issues (e.g., trying to cure Covid-19 with household cleaners) can be ruled out rather than advancing to the later stages of decision-making where they get discussed and debated.

Encouragingly, the research on rating creative products suggests that individuals can be trained to improve their accuracy in recognizing and distinguishing creative products. There is some indication in the literature that novices, and certainly quasi-experts, can be brought closer to expert judgments with some effort. Such methods to help non-expert judges include showing them already-rated creative work to help them calibrate their own judgements (Dollinger & Shafran, 2005), having an expert offer dynamic feedback (Storme, Myszkowski, Celik, & Lubart, 2014), and having them rate the creative work on many different dimensions (such as surprise or gracefulness) that are related to creativity before assigning a more holistic score (Besemer, 1998; Cropley & Kaufman, 2012). Might such steps be more effective for distinguishing either particularly creative or uncreative work? How might individual predictors, from specific creative abilities to personality to intelligence to demographic variables impact the benefit of such training or additional guidance? The more we can learn how to best tailor interventions to improve ratings for everybody, the easier (or cheaper) it may be to make quality judgments of creative work.

CRediT authorship contribution statement

Steven E. Stemler: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing. James C. Kaufman: Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013.

Amabile, T. M. (1996). *Creativity in context: Update to The Social Psychology of Creativity.* Boulder, CO: Westview Press.

Baer, J. (1998). The case for domain specificity in creativity. *Creativity Research Journal, 11*, 173–177.

Baer, J., & Kaufman, J. C. (2008). Gender differences in creativity. *Journal of Creative Behavior, 42*, 75–106.

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., Kaufman, J. C., & Riggs, M. (2009). Rater-domain interactions in the Consensual Assessment Technique. *International Journal of Creativity and Problem Solving, 19*, 87–92.

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., et al. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity, 21*, 75–84.

Besemer, S. P. (1998). Creative product analysis matrix: testing the model structure and a comparison among products-three novel chairs. *Creativity Research Journal, 11*(4), 333–346.

Birney, D. P., Beckmann, J. F., & Sebah, Y. Z. (2016). More than the eye of the beholder: The interplay of person, task, and situation factors in evaluative judgments of creativity. *Learning and Individual Differences, 51*, 400–408.

Caroff, X., & Besançon, M. (2008). Variability of creativity judgments. *Learning and individual differences, 18*, 367–371.

Cattell, J., Glascock, J., & Washburn, M. F. (1918). Experiments on a possible test of aesthetic judgment of pictures. *The American Journal of Psychology, 333–336.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. *Gifted Child Quarterly, 49*, 283–291.

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: empirical validation of the Creative Solution Diagnosis Scale (CSDS). *Journal of Creative Behavior, 46*, 119–137.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*, 35–40.

Dollinger, S. J., & Shafran, M. (2005). Note on the Consensual Assessment Technique in creativity research. *Perceptual and Motor Skills, 100*, 592–598.

Elliott, J. G., Stemler, S. E., Grigorenko, E. L., Sternberg, R. J., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal, 37*(1), 83–103.

Ericsson, K., & Simon, H. (1980). "Verbal reports as data". *May Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215.
