J. Intell. **2015**, *3*, 73-90; doi:10.3390/jintelligence3030073

OPEN ACCESS
Journal of
Intelligence
ISSN 2079-3200
www.mdpi.com/journal/jintelligence

Article

Differences in Judgments of Creativity: How Do Academic Domain, Personality, and Self-Reported Creativity Influence Novice Judges' Evaluations of Creative Productions?

Mei Tan [1], Catalina Mourgues [1], Sascha Hein [2], John MacCormick [3], Baptiste Barbot [1,4] and Elena Grigorenko [1,5,*]

[1] Yale Child Study Center, 230 South Frontage Road, New Haven, CT 06520, USA; E-Mails: mei.tan@yale.edu (M.T.); catalina.mourgues@yale.edu (C.M.); bbarbot@pace.edu (B.B)
[2] Department of Psychological, Health, and Learning Sciences, University of Houston, 3657 Cullen Boulevard, Houston, TX 77204-5029, USA; E-Mail: sdhein@uh.edu
[3] Department of Computer Science, Dickinson College, 28 North College Street, Carlisle, PA 17013-2737, USA; E-Mail: jmac@dickinson.edu
[4] Psychology Department, Pace University, One Pace Place, New York, NY 10038, USA
[5] Moscow State University of Psychology and Education, ul. Sretenka, 29, Moscow 127051, Russia

***** Author to whom correspondence should be addressed; E-Mail: elena.grigorenko@yale.edu; Tel.: +1-203-737-2316.

Academic Editor: Robert J. Sternberg

Received: 1 July 2015 / Accepted: 9 September 2015 / Published: 14 September 2015

Abstract: Intelligence assessment is often viewed as a narrow and ever-narrowing field, defined (as per IQ) by the measurement of finely distinguished cognitive processes. It is instructive, however, to remember that other, broader conceptions of intelligence exist and might usefully be considered for a comprehensive assessment of intellectual functioning. This article invokes a more holistic, systems theory of intelligence—the theory of successful intelligence—and examines the possibility of including in intelligence assessment a similarly holistic measure of creativity. The time and costs of production-based assessments of creativity are generally considered prohibitive. Such barriers may be mitigated by applying the consensual assessment technique using novice raters. To investigate further this possibility, we explored the question: how much do demographic factors such as age and gender and psychological factors such as domain-specific expertise, personality or self-perceived creativity affect novices' unidimensional ratings of creativity? Fifty-one novice judges from

three undergraduate programs, majoring in three disparate expertise domains (*i.e.*, visual art, psychology and computer science) rated 40 child-generated Lego creatures for creativity. Results showed no differences in creativity ratings based on the expertise domains of the judges. However, judges' personality and self-perception of their own everyday creativity appeared to influence the way they scored the creatures for creativity.

Keywords: creativity; consensual assessment technique; novice judges

1. Introduction

The assessment of intelligence has long been dominated by psychometric theories (e.g., as synthesized by the Cattell-Horn-Carroll [CHC] theory [1–3]) that define intelligence as a network of highly specific mental processes. IQ measurement has thus evolved into a precision exercise, employing a collection of tests that focus on component cognitive processes and are validated via a factor-analytic approach. Ironically, IQ tests have their roots in the work of Alfred Binet, who favored more qualitative rather than quantitative views of intelligence, and whose first intelligence battery included tests for "imagination" [4,5]. So when we are exhorted to "resist being blinded by the landmark importance of the current CHC taxonomy" [6], we agree that other avenues should be explored.

Here, we choose to put forth instead a broader "systems" view of intelligence as a basis for assessment, represented by the correspondingly broad view (*i.e.*, beyond divergent thinking or ideational fluency) of creativity that this systems view includes. The theory of successful intelligence [7] proposes that intelligence may be conceived as a collection of cognitive skills that could be characterized as analytical, practical or creative. In this conception, creativity is an integral part of intellectual functioning. The importance of considering broader theories of intelligence lies in whether new and different assessments of intelligence can do more than current instruments within the sphere of education, where the stakes for success and failure are high, and where understanding strengths and weaknesses may be more important and meaningful to educators than clinical diagnoses based on evermore fine-grained processing skills. It has in fact been argued that holistic views of human intelligence may be more useful for understanding of human achievement and success and may have more applications within the workplace and schools [8]. In this article, we hope to contribute to the discussion of the global importance of intelligence and creativity by proposing, first, that an assessment of a broad conception of intelligence—one that includes creativity—may play a useful role in educational settings. Second, that methods for the assessment of creativity that may be implemented as a part of educational practice should be explored continuously, even in the face of arguments of their technical and practical costs.

According to Sternberg's theory of successful intelligence, analytical, practical and creative skills are three key areas of skill that an individual draws upon to adapt to, change, or shape her environment, to best employ strengths and compensate for weaknesses in the pursuit of personally important goals [7,9]. Within this framework of intelligence, creativity encompasses the skills that a person needs to make original things, imagine, suppose, and cope with novel situations [7,10,11]. As part of an individual's cognitive and educational profile, how creativity might be defined and assessed matters.

In our work developing an assessment battery based upon Sternberg's theory of successful intelligence [12–14], an observation schedule was proposed in addition to several group-administered paper and pencil activities. Both formats include subtests for creativity, yet the observation schedule offers the child an opportunity to produce an original object in an activity that is distinct from any academic task, such as writing. The question then arises, *how* should these productions be scored? Equally important is the question of *who* should score such productions?

Here we add to this literature by deliberately employing two approaches: first, we used the consensual assessment technique (CAT) [15,16] in its original form, which simply asks judges to consider whether a product is creative without assessing any other dimensions. The most important facet of the CAT (described in more detail later) is that a product may be deemed "creative" if a set of "appropriate observers" agree that it is [15], thus relying wholly on implicitly held subjective criteria. Second, although the CAT specifies the use of experts—or at least those familiar with the domain—to rate creative products, we asked novice raters to carry out these evaluations. We then evaluated the fitness and appropriateness of our methods by looking at how individual differences between novice judges may affect their rating behavior when evaluating creativity using the consensual assessment technique. This is important because rater characteristics may lead to an invalid evaluation of creativity, so understanding the sources of influence on raters' scores may be of interest when seeking valid scores of creativity. We examined particularly raters' severity/leniency—general tendency to score too harshly or too leniently resulting in lower or higher scores than expected [17]; their discrimination—how well each rater is able to distinguish the finer incremental levels of creativity as represented in the creative productions [18]; and their consistency—ability to produce systematic ratings that reflect some shared variance or common conception of the target construct [19].

Specifically, we explored how these three parameters of rater behavior—severity, discrimination, and consistency—in samples of novice judges, are associated with several variables: the judges' demographic characteristics, domains of developing expertise (*i.e.*, visual art, psychology and computer science), personality traits, and self-assessed creative behaviors. That is, are judges' rating characteristics (*i.e.*, severity, discrimination, consistency) related to these demographic and individual dimensions? And are any areas of judges' creative behavior (as measured by a self-assessment measure) related to the characteristics of their creativity judgments? We expected that the domains of developing expertise (based on the use of experts in the CAT) and personality traits of the judges, as explored in other studies [20,21], might be associated with their rating characteristics. In addition, we expected that judges who score higher in their self-assessment of creative behaviors would be more severe and consistent in their rating of creative products, as more creative judges have been found to be more severe and reliable [19].

The use of the CAT and the characterization of novice rater behavior is important if we want to consider a view of intelligence that includes a broadly defined creativity, the accurate assessment of which may play a useful role in education by helping define students' optimal roads to success. Before presenting our study, we briefly review literature on the CAT, creativity studies employing both experts and novices, and studies examining rater differences in the evaluation of creativity.

2. The Consensual Assessment Technique

The theory of successful intelligence defines creativity as the capacity to generate new ideas, create and design in activities like writing, drawing, building and imaginative play. It may be discerned particularly well using problems assessing how an individual copes with relative novelty [7,9]. Most definitions of creativity agree on two necessary elements—novelty and appropriateness [22–24]. However, applying these two criteria to a creative product involves the subjective evaluation of these elements as they are expressed within a given context or domain of expertise, such as music, poetry, or advertising. So an important question arises: Who should judge? Within the field of research on creativity, the CAT [15] has become a well-known, often-used method for scoring creative products. The assessment of creative productions, while perhaps more commonly applied to eminent creators, was developed for and is particularly useful in the study of everyday creativity (little c), according to Hennessey and Amabile [25], because making a creative product is a holistic task that engages all of the person-level resources important for creativity [26]. In addition, product-based assessment is supported developmentally: in childhood, creativity and its components are only moderately differentiated [27]. Therefore a holistic approach is not only warranted but probably more efficient than resource-based approaches (e.g., divergent thinking) to creativity assessment. The CAT is based on the assertion that those most knowledgeable about the context or domain in which a creative product is made—experts—will agree on what is creative within that domain. The agreement or consensus among experts, in other words, may define and validate what is considered creative in their domain of expertise. Inter-rater reliabilities among expert judges using the CAT have been found to be generally high, typically in the 0.70–0.90 range [16,28–30]. Yet, given the difficulty and expense of securing domain experts, a second question arises: How reliable are the ratings of novice judges [31], and what variables might influence this reliability?

Several studies comparing different groups of raters, from experts to novices to peers, have been carried out focusing on different types of creative products. We draw here very purposefully only from those studies using the CAT where creativity is scored as a unidimensional continuum, following the original method designed by Amabile, which has been used successfully in this form over three decades. We do so first because of this approach's attractive simplicity; for group assessments that may generate a mass of productions to score, simpler, less time-consuming approaches are desirable. A second reason is the uncertain meaning of multidimensional ratings of creativity (e.g., scoring separately for originality and appropriateness), in comparison to a unidimensional approach [15,32]. Although several studies have looked at the consistency of experts *vs.* novices focusing on the use of specifically scaled tools for rating creativity (e.g., scoring originality and appropriateness separately) [33–35], such rubric-assisted judgments of creative products may be limited as they are generally formulated upon fairly specific theories or conceptions of creativity [29,36] and results may be skewed by the semantic interpretations made by raters unless training is undertaken [37]. In fact, Runco and Charles [38] found differential relationships of ratings of originality and appropriateness to judgments of creativity, and found that judgments of originality were significantly more accurate than judgments of appropriateness. A subsequent study by Caroff and Besançon [39] revealed that raters integrate originality and appropriateness in particular ways that show varied relationships with mean creativity scores [39]. Thus, the CAT may be considered more easily and broadly usable since it calls for a unidimensional

consideration of creativity and requires no training [29]. The CAT may also be considered a better format for novice raters, whose rating behaviors may generalize across dimensions of creativity, making a unidimensional rating the most apt for the task at hand as well as for the focus of this study, which is the influence of various factors on the scoring behavior of novice raters. We begin our consideration of the usefulness of novice raters by examining some expert-novice comparisons using the CAT.

In a study on the creativity of musical composition using three different groups of judges, Priest [40] reported that when 21 non-music majors, 22 elementary music specialists, and 22 instrumental teachers used the CAT to rate the creativity of five undergraduate student compositions, listening to an audio recording, the student and teacher groups each displayed similarly high reliability when adjusted for equal numbers of raters per group (ICC = 0.89−0.96). The students and instrumental teachers also agreed on their ranking of the five compositions. Other studies, however, have reported contrasting findings. Kaufman and colleagues [19] asked 10 published poets (experts) and 106 college undergraduates (from various ethnic backgrounds; academic majors unknown) to use the CAT to rate 204 poems generated on-line by college undergraduates who participated for class extra credit. Results showed not only that experts rated the poems more severely (lower group mean across poems), their reliability, when adjusted for different numbers of raters, far exceeded that of the student ratings (Cronbach's $\alpha = 0.804$ *vs.* 0.575). This, the authors proposed, may reflect the internalized standards inevitably established by experts through experience [19]. A later study also suggested that expert-novice agreement may depend upon the ubiquity of the activity (poetry *vs.* music, for example) amongst the general population, with more popular or common tasks being easier for experts and novices to agree on [41].

Similar differences were found when creativity ratings of expert movie critics (extracted from a database of compiled published reviews) were compared with individuals who regularly reviewed movies on-line via boxofficemom.com and imdb.com, and with 169 university students (ethnically diverse; academic major unknown) [42]. Again, professional critics were found to be more severe raters (tended to rate movies more stringently); reported correlations between students and experts' ratings were only moderate ($r = 0.43$, $p < 0.01$). On-line novice critics' ratings correlated more closely with the professionals ($r = 0.72$ and $r = 0.77$, $ps < 0.01$, for the two on-line sites), which suggest that there may be an intermediate level of rater, a quasi-expert, between the novice and expert.

Yet, whatever the level of expertise of judges, rater differences may be found to influence creativity scores. These differences can arise from such large, complex constructs as cultural differences [43,44]. For example, Niu and Sternberg [44] asked Chinese and American novice judges (nine psychology graduate students from each country) to evaluate the creativity of student-produced collages and drawings. The Chinese judges showed higher reliability or more consensus; American judges as a group were more severe. These differences were attributed to the different ways in which each culture—one collective, one individualistic—values and conceives creativity. Judges who have been assessed as being original themselves were found to be more severe in one study [45], and more discriminating with respect to the originality of advertisements in another [39]. Personality differences have also been detected. It was found [21] that those novice raters who had a preference for novelty as assessed using Pearson's Novelty Experiencing Scale (NES) [46], linked with Openness from the Big Five Model [47] tended to give greater weight to originality in their conceptions of creativity.

A few studies accessing only non-expert judges (undergraduate university students) that have used the CAT have found these judges to show reliable scoring behaviors [32,48]. Researchers [48], for

example, asked college undergraduates from an introductory psychology course to rate the creativity of drawings of alien creatures. The overall inter-rater reliability of the creature ratings, based on Rasch analysis (for a coefficient similar to Cronbach's α), was 0.83. In these studies, the novice raters were neither characterized by personality traits, nor creative ability. Notably, all studies using undergraduate novice raters have accessed psychology students or students of unknown (unreported) majors. In the present study, we access a sample of undergraduate students majoring in three distinct subjects to examine possible sources of rater difference in novice judges' ratings of creativity using of a set of 40 Lego creatures produced by children 9 to 12 years of age from a variety of educational settings. The task is part of an assessment battery for intelligence that is under development. Legos were selected as a common construction tool, like Crayons—familiar to all children, easy to wield, and conducive to varying degrees of complex productions. It was not expected that experts in the field of Lego-building would be needed to assess these creatures for creativity.

3. Method

3.1. Participants

The 51 participants in this study were undergraduate students who attended three different universities. Students of one particular academic major were recruited from each university: Fine Arts (29.4%, $n = 15$), Psychology (27.5%, $n = 14$), and Computer Science (43.1%, $n = 22$). Participants were 49.0% female. However, the gender distribution within the three domains of expertise was not homogeneous: 78.6% of the psychology students ($n = 11$), 60% of the fine arts students ($n = 9$), and 22.7% ($n = 5$) of the computer science students were female. Within the total sample, ages ranged from 18 to 35 years ($M = 21.31$, $SD = 2.58$). Most students were white/Caucasian (60.8%); the rest identified as Asian (19.6%), and Hispanic, African American or other (19.7%). Table 1 shows the demographics for each school/domain of expertise.

Table 1. Demographics, self-assessed creativity and personality variables of the judges.

	Psychology (*n* = 14)				**Arts (*n* = 15)**				**Computer Science (*n* = 22)**				**Total Sample (*n* = 51)**				
	Min	**Max**	***M***	***SD***	**Min**	**Max**	***M***	***SD***	**Min**	**Max**	***M***	***SD***	***M***	***SD***	***skewness***	***kurtosis***	n^2p
Age	18	22	21.07	1.07	19	35	23	4.16	19	21	20.32	0.72	21.31	2.58	3.31	15.44	0.23
GPA	3.20	4.00	3.72	0.20	2.75	4.00	3.53	0.42	3.00	4.00	3.62	0.32	3.62	0.33	−0.96	−0.09	0.05
Creativity − Creatures	2.28	6.05	4.15	0.79	2.65	4.53	3.78	0.54	2.73	5.30	3.84	0.58	3.91	0.64	0.16	2.54	0.02
K-D Total	2.26	4.46	3.10	0.53	2.96	3.80	3.42	0.24	2.24	4.44	3.22	0.54	3.25	0.48	0.12	0.67	0.07
K-D Self/Everyday	2.73	4.55	3.77	0.55	3.00	4.91	3.74	0.54	2.09	4.36	3.57	0.57	3.67	0.55	−0.49	0.64	0.03
K-D Scholarly	2.27	4.45	3.44	0.58	2.18	4.45	3.42	0.65	2.09	4.55	3.35	0.65	3.40	0.62	−0.24	−0.60	0.00
K-D Performance	1.00	4.60	2.50	1.12	1.70	4.50	2.94	0.69	1.00	4.20	2.77	0.92	2.75	0.92	−0.15	−0.62	0.03
K-D MechSci	1.33	4.11	2.35	0.78	1.67	3.78	2.64	0.57	2.00	4.78	3.57	0.73	2.96	0.88	0.13	−0.68	0.39
K-D Artistic	1.67	4.67	3.27	0.92	3.56	5.00	4.34	0.44	1.56	4.44	2.77	0.86	3	1	−0.26	−1.09	0.43
BFI Agree	22	43	36.64	5.62	22	45	37.40	6.48	19	40	31.27	5.47	34.55	6.40	−0.46	−0.38	0.05
BFI Con	20	44	31.21	7.45	18	43	32.80	6.88	18	43	31.18	5.66	31.67	6.46	−0.20	−0.47	0.21
BFI Extra	14	31	24.86	4.87	8	36	22.33	7.26	10	40	22.50	7.30	23.10	6.67	0.11	0.34	0.01
BFI Neuro	16	31	23.57	4.20	13	36	24.27	6.57	13	35	23.36	5.84	23.69	5.58	0.21	−0.21	0.03
BFI Openness	29	46	38.07	3.97	35	48	41.93	3.61	24	43	32.95	4.82	37.00	5.69	−0.29	−0.22	0.01

K-D Total = K-DOCS total self-assessed creativity; K-D Everyday = K-DOCS everyday creativity; K-D Scholarly = K-DOCS scholarly creativity; K-D Performance = K-DOCS performance creativity; K-D MechSci = K-DOCS mechanical/scientific creativity; K-D Artistic = K-DOCS artistic creativity; BFI Agree = agreeableness; BFI Con = conscientiousness; BFI Extra = extraversion; BFI Neuro = neuroticism; BFI Openness = openness; n^2p = univariate partial eta squared, generated using MANOVAs testing for group differences in the respective outcomes. Scores on the K-DOCS (K-D) reflect the mean across all items. Scores on the BFI reflect sum scores.

3.2. Measures

Creative Productions to Be Rated

For this study, the judges rated 40 Lego-creatures selected from a larger sample of creatures collected in a variety of educational settings—public schools, private schools, and summer programs—from boys and girls 9–12 years old. The students who made the creatures were given a special set of Legos (40 pieces) from which all pieces suggestive of a creature were removed (e.g., bodies, heads, eyes), and which contained a few unusual pieces (e.g., magenta disc, orange bricks). The instructions were: "Use these Legos to make a creature. It can be anything! You have 10 min and you can use whichever pieces you like and as many pieces as you like." The resulting Lego sculptures were then photographed. The creatures in the study described here were selected for clarity of the photograph and for variety. Of three typical categories of types of creature generally produced—building-like (e.g., likes houses or towers), robot-like (e.g., bi-pedal with distinct head and arms), or animal-like (e.g., quadrupeds, birds, snakes, or mythical creatures)—13, 14, and 13 representatives from each category, respectively, were selected. Students rated all 40 Lego creatures (randomized for each participant) on a scale of 1 (*not creative*) to 7 (*very creative*). They were instructed to try to use the complete scale (1–7), comparing the 40 given creatures to each other, and basing their scores on their own ideas of what is more or less creative.

3.3. Personality

The Big Five Inventory (BFI) [49,50], a 44-item inventory, was used to measure individuals on the Big Five Factors/dimensions of personality—Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. It asks respondents to rate their agreement from 1 (*disagree strongly*) to 5 (*agree strongly*) with statements such as "I am someone who is outgoing, sociable" or "I am someone who does a thorough job"; Extraversion (8 items), Cronbach's $\alpha = 0.847$; Agreeableness (9 items), Cronbach's $\alpha = 0.843$; Conscientiousness (9 items), Cronbach's $\alpha = 0.847$; Neuroticism (8 items), Cronbach's $\alpha = 0.742$; and Openness (10 items), Cronbach's $\alpha = 0.775$.

3.4. Self-Perceptions of Creativity

The Kaufman Domains of Creativity Scale (K-DOCS) [51] is a 50-item rating scale that asks respondents to rate themselves comparatively with their peers on creativity from 1 (*much less creative*) to 5 (*much more creative*). The items span five proposed domains of creativity: Self/Everyday, Scholarly, Performance (encompassing writing and music), Mechanical/Scientific, and Artistic. Higher scores indicate higher self-perceived creativity. Cronbach's α for the total scale (50 items) was 0.883; Self/Everyday (11 items), Cronbach's $\alpha = 0.769$; Scholarly (11 items), Cronbach's $\alpha = 0.795$; Performance (10 items), Cronbach's $\alpha = 0.863$; Mechanical/Scientific (9 items), Cronbach's $\alpha = 0.830$; and Artistic (9 items), Cronbach's $\alpha = 0.900$.

4. Procedure

Students were recruited within their departments through an e-mail invitation. Interested students accessed the study materials through an online web-survey site (Qualtrics). Initial pages of the site

informed them of the purpose of the study, described the activities they would be asked to complete, and explained that their participation should be voluntary and would remain confidential. Students were then asked to give their consent to participate in the study. Basic demographic information was collected, including gender, age, class/year at university, academic major, ethnicity, and state of residence. Parents' levels of education and occupation were also collected as a proxy for socio-economic status. Students then rated 40 Lego creatures (randomized for each participant), and completed the BFI and the K-DOCS rating scales, in that order.

5. Data Analysis

First, the judges in the three domains were compared on their demographic variables, as well as their personality and creativity ratings. Next, we examined correlations between the judges' creativity ratings of the creatures, and the BFI and K-DOCS subscales. We also analyzed gender differences. Next, we examined the judges' rating behaviors. The creativity ratings of the Lego creatures were analyzed using a Many-Facets Rasch Measurement (MFRM) approach using Facets software [52]. This is as an alternative to the classic average composite scoring of the CAT; it is used to ascertain fine-grained differences in the assessment of creative products (e.g., pieces of writing, paintings, musical composition) by multiple raters, providing information not only on productions but also on rater behaviors, while accommodating small sample sizes very well. MFRM converts qualitative observations into linear quantities—logit scores—by modeling responses as functions of item (creature) and person (rater) characteristics. In addition to providing estimates of the relative consistency of ratings and the degree of agreement between judges, this approach allows the evaluation of the extent to which each judge is using the scoring rubric in a manner that is internally consistent.

Thus, using MFRM, we estimated the judges' ratings as logit scores (*i.e.*, the probability of each creature's rating based on that creature's overall level of creativeness within the set, and the severity of each rater with respect to the other raters) and compared judges' scoring behavior by domain of expertise. We then considered three aspects of the judges' scores. First, we looked at judges' leniency/severity, that is, whether a judge tended to score creatures as more creative (scoring leniently, using the higher end of the scale more) or less creative (scoring severely, using the lower end of the scale more). We examined judge leniency/severity using the mean of the logit scores of each judge as well as the mean of their raw scores. Second, we looked at the judges' ability to discriminate—to recognize various degrees of creativity. Logit score standard deviations higher than 1 indicated a judge's ability to discriminate creatures' creativity at a higher level than expected; values lower than 1 indicated a judge's ability to discriminate less than expected. And third, we looked at judges' consistency, or how much a judge tended to agree with all of the other judges as a group. To do this, we used point biserial correlations (correlating one judge's scores against the group scores of the rest of the judges) using the logit scores and Cronbach's α, obtained using the raw scores. Raters' leniency/severity, discrimination, and consistency were compared across domains of expertise.

Next, to see whether any of our variables were related to the leniency/severity, discrimination, and consistency of our raters, we used a median score to split each group to create equal high and low groups in each category (*i.e.*, a highly severe group and highly lenient group; a highly discriminating group and a less discriminating group; and a highly consistent group and a less consistent group). These high *vs.*

low groups were then compared with respect to the personality variables (BFI) and self-reported creativity (K-DOCS) variables. This allowed us to explore the relationships between rater personality, rater creativity and their rating behavior using multivariate analyses of covariance (MANCOVAs), controlled for gender. For categorical variables, chi-square tests were used to investigate the association between the judgment groups (high *vs.* low groups severity, consistency and discrimination) and categorical predictors (e.g., gender). For continuous variables, a correlational analysis was conducted using Spearman's rank correlations (more robust in the context of small samples sizes/non-normal data).

Finally, linear regression analyses using successive stepwise models were conducted to see which personality (BFI factors) and creativity variables (K-DOCS factors) predicted judges' severity, consistency and discrimination. Due to the limited sample size, only four variables were included as independent variables in each of the three regression models: two from the K-DOCS—everyday creativity and artistic creativity; and two from the BFI—agreeableness and conscientiousness. These variables were selected considering previous studies in the literature [21] and our preliminary analyses.

6. Results

6.1. Description of Judges, Domain of Expertise

The judges from the three domains were compared across all of the variables. Of the demographic variables, differences across the groups were observed only for age ($F_{(2,48)} = 6.053$, $p < 0.005$): artist judges were older than the computer science judges ($p = 0.003$).

With respect to personality, significant group differences were found in two scales of the BFI, agreeableness ($F_{(2,48)} = 6.187$, $p = 0.004$) and openness ($F_{(2,48)} = 20.334$, $p < 0.001$): computer science judges obtained lower scores in both agreeableness and openness than the art ($p = 0.009$, $p < 0.001$) and psychology students ($p = 0.029$, $p = 0.003$). Regarding the K-DOCS, differences were found in the mechanic/scientific creativity ($F_{(2,48)} = 15.071$, $p < 0.001$) and artistic creativity scales ($F_{(2,48)} = 18.291$, $p < 0.001$): computer science students scored higher than the art and psychology students in mechanic/scientific creativity ($ps < 0.005$). Art students scored higher in self-assessed artistic creativity than both the computer science and psychology students ($ps < 0.005$). Table 1 shows the descriptive statistics for the judges' personality dimensions and self-perceptions of creativity by domain, including the univariate effect sizes (partial eta squared) for each variable.

6.2. Correlations and Gender Differences

Correlation analysis showed a positive relationship between judges' creativity scores of the creatures and the judges' everyday creativity (K-DOCS) and agreeableness (BFI). The everyday creativity scale of the K-DOCS was also positively correlated with agreeableness, conscientiousness and extraversion, and negatively related to neuroticism. Table 2 shows the bivariate correlations for all of the variables. MANCOVAs for all of the scales of the K-DOCS, controlling for academic domains, showed no differences by gender. Of the BFI subscales, only conscientiousness ($F_{(1,50)} = 5.501$, $p < 0.023$) indicated a difference by gender, with females being more conscientious than males (males, $M = 29.84$, $SD = 6.31$; females, $M = 33.5$, $SD = 6.17$).

Table 2. Spearman's rank-order correlations between age, gender, creativity scoring and self-assessed creativity and personality.

	Age	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1. Gender	0.176	1															
2. M Logit	−0.052	0.240	1														
3. SD Logit	0.065	−0.189	−0.264	1													
4. Consistency	−0.146	0.150	0.293 *	−0.072	1												
5. Discrimination	−0.065	0.050	0.343 *	−0.023	0.729 **	1											
6. K-D Total	0.096	−0.017	0.281 *	0.225	0.022	0.200	1										
7. K-D Everyday	0.282 *	0.213	0.393 *	−0.105	0.360 *	0.326	0.570 **	1									
8. K-D Scholarly	0.125	−0.054	0.022	0.267	−0.049	0.004	0.508 **	0.314 *	1								
9. K-D Perform	−0.039	−0.111	0.129	0.194	−0.127	0.091	0.789 **	0.254	0.150	1							
10. K-D MechSci	−0.303 *	−0.349 *	0.137	0.096	−0.019	0.089	0.456 *	0.056	0.093	0.306 *							
11. K-D Artistic	0.272	0.268	0.206	0.180	−0.001	0.133	0.657 **	0.281	0.136	0.502 **	−0.094	1					
12. BFI Agree	0.407 *	0.311 *	0.351 *	−0.287 *	0.193	0.194	0.118	0.506 **	−0.062	−0.051	−0.209	0.249	1				
13. BFI Con	0.219	0.290 *	−0.056	−0.263	−0.040	−0.195	0.049	0.338 *	−0.051	−0.065	0.060	−0.047	0.265	1			
14. BFI Extra	−0.004	0.086	0.070	0.074	−0.128	−0.083	0.204	0.310 *	0.114	0.104	0.058	0.083	0.056	0.019	1		
15. BFI Neuro	−0.013	0.027	0.028	0.080	0.057	0.197	−0.216	−0.309 *	0.024	−0.143	−0.189	−0.066	−0.229	−0.422 *	−0.221	1	
16. BFI Openness	0.284 *	0.028	0.137	0.093	0.021	0.058	0.383	0.285 *	0.252	0.284 *	−0.347 *	0.637 **	0.365 *	−0.040	0.211	−0.070	1

M Logit = Mean creativity logit score; SD Logit = Standard deviation of the creativity logit score; K-D Total = K-DOCS total self-assessed creativity; K-D Everyday = K-DOCS everyday creativity; K-D Scholarly = K-DOCS scholarly creativity; K-D Perform = K-DOCS performance creativity; K-D MechSci = K-DOCS mechanical/scientific creativity; K-D Artistic = K-DOCS artistic creativity; BFI Agree = agreeableness; BFI Con = conscientiousness; BFI Extra = extraversion; BFI Neuro = neuroticism; BFI Openness = openness. * $p < 0.05$, ** $p < 0.001$.

6.3. Judges' Rating Behavior, by Domain of Expertise

No between-group differences were found in the severity (raw and logit scores used), discrimination (logit scores used), or consistency (point biserial correlations used) of the judges' scoring with regard to the domain of expertise. The correlations between arts students and psychology students (using raw scores) was $r = 0.841$, between psychology and computer science $r = 0.900$, and between art and computer science $r = 0.865$ ($ps < 0.001$). Fisher's *z*-test showed that the correlations were not significantly different between the groups.

This result is confirmed by the logit scores obtained from the MFRM analysis, which suggested a low variability amongst the judges in how each group used the creativity rating scale. That is, the logit scores varied between +1 and −1 logits, with the exception of two raters, one more lenient (1.72; psychology) and the other more severe (−1.35; psychology), indicating that generally, all of the judges used the scale similarly. More importantly, all of the judges across domains similarly used the complete scale (see Table 3), producing scores with fairly normal distributions.

Table 3. Frequency of ratings by judges' domain of expertise.

	Psychology		**Arts**		**Computer Science**	
Scale	Fq	%	Fq	%	Fq	%
1	33	5.9	37	6.2	56	6.4
2	67	12.0	85	14.2	137	15.6
3	94	16.8	149	24.8	176	20.0
4	120	21.4	142	23.7	198	22.5
5	117	20.9	104	17.3	173	19.7
6	91	16.3	54	9.0	117	13.3
7	38	6.8	29	4.8	23	2.6

Fq = number of ratings.

6.4. Severity, Discrimination and Consistency

When the judges were split using the median logit score (median = 0.07) to make severe and lenient groups, subsequent ANOVAs showed that the severe and lenient judges differed only in their scores on everyday creativity ($M_{Lenient} = 3.8$, $SD_{Lenient} = 0.42$; $M_{Severe} = 3.4$, $SD_{Severe} = 0.60$; $F_{(1,50)} = 4.382$, Cohen's $d = 0.745$, $p \leq 0.012$) and agreeableness ($M_{Lenient} = 36.4$, $SD_{Lenient} = 4.60$; $M_{Severe} = 32.8$, $SD_{Severe} = 7.3$; $F_{(1,50)} = 1.465$, Cohen's $d = 0.598$, $p = 0.042$). No association between severity groups and domain of expertise were found. Linear regression analyses were conducted to explore which of the individual-level variables (personality, everyday creativity) was the best predictor of the severity of the judges. Only everyday creativity was significant in the model ($R^2 = 0.236$, $R^2_{adj} = 0.170$, $\beta = 0.345$, $t = 2.178$, $p = 0.007$). Finally, when gender differences were explored, significant differences were found ($\chi^2 = 4.404$, $p = 0.036$), revealing that there were more males in the severe group than females.

A discrimination score, the standard deviation of the logit, reflecting the ability of a judge to distinguish between various levels of creativity among the creatures, was obtained for each judge. The judges were split into two groups, those who discriminated more than expected and those who discriminated less than expected (median = 1.22). The two groups were compared on all variables. No

significant differences were found (the largest effect size was found for Conscientiousness, Cohen's $d = 0.524$). According to the linear regression analysis, the consistency of the judges was significantly predicted only by the judges' everyday creativity ($R^2 = 0.406$, $R^2_{adj} = 0.148$, $\beta = 0.406$, $t = 3.114$, $p = 0.003$).

The consistency of each judge was initially assessed by generating point biserial correlations (correlations between the score of one judge and the group score of the rest of the judges). The judges were then split into two groups (median = 0.586), those with higher correlations with the group, those with lower. No group differences were observed in the subsequent ANOVA for any of the variables of the study (the largest effect size was found for Performance, Cohen's $d = 0.357$). In an additional analysis for consistency, we used Cronbach's α, the most traditional measure of rater consistency in the context of CAT, to explore consistency differences by domain groups. The psychology students showed an $\alpha = 0.889$, the art students $\alpha = 0.836$, and for computer science students $\alpha = 0.893$. Rater consistency for the total group was $\alpha = 0.873$. Application of the Spearman-Brown formula indicated that for a group of judges from each domain to reach an acceptable Cronbach's α of 0.70, one would need at least 4 judges from psychology, 6 judges from computer science, and 7 in art. Correspondingly, to adjust for the numbers of raters, which could potentially inflate the α, we also computed the single measure intraclass correlations (ICC) coefficient (reflecting the average inter-correlation between judges in each domain): psychology students, ICC = 0.364; art students, ICC = 0.254; and computer science students, ICC = 0.274. These ICCs reflect that the reliability of each judge within each domain was rather low, with psychology students seeming the more consistent, *i.e.*, in higher agreement with each other. Finally, the linear regression analysis showed that levels of everyday creativity ($\beta = 0.443$, $t = 3.249$, $p = 0.002$) positively predicted scores of discrimination, while levels of conscientiousness ($\beta = -0.345$, $t = -2.534$, $p = 0.015$) negatively predicted discrimination ($R^2 = 0.460$, $R^2_{adj} = 0.179$).

7. Discussion

According to our results, novice judges from different domains did display some differences in personality and self-assessed creativity, which are associated with the main characteristics of creativity rating. Judges concentrating in the field of computer science scored lower on agreeableness and openness than judges concentrating in art and psychology. In addition, judges in computer science scored higher on mechanic/scientific creativity than art and psychology judges; judges majoring in art scored higher on artistic creativity than judges in computer science and psychology. Here, we need to note that the computer science group included the majority of males, therefore gender maybe a confounding factor in these results. Additionally, each major came from a different school, introducing the possibility of school related differences in these results.

What was unexpected was that despite these differences, judges majoring in different domains (even coming from different schools and representing gender differences) did not score the creatures for creativity significantly differently as groups. While conceivable that novice judges studying art—who might have ostensibly scored as knowledgeable novices in visual/visual-spatial productions [42,53]—might score with greater severity (*i.e.*, more like "experts"), as described by Kaufman and colleagues [19], they did not. And although novice judges in psychology seemed to score with slightly more reliability and consistency—the differences between rater groups remain inconclusive. That is, domains of expertise

had no significant distinguishing effect on these novice judges' scoring behavior. This might possibly be attributable to the ubiquity of Legos (the medium of the creation), as suggested in the expert-novice study by Kaufman and colleagues [41]. Novice judges (*i.e.*, university students) of unknown academic major have been used as raters before [19,48]; domain effects were not included in those studies. Concerning the consistency of novices' ratings, the point we make here is that in larger numbers, novices (and others) can agree on a unidimensional scale of creativity, as indicated in our analyses of judges' consistency. How many judges are needed is important for practical purposes, as the need for 1–2 expert judges or 20 novice judges would be an important consideration in a situation in which 50–60 productions need to be scored.

Of importance also is that other personal factors such as dimensions of personality and self-assessed creativity did appear to play a role in these judges' scoring. Specifically, those judges who scored higher in agreeableness according to the BFI tended to be more lenient. This is not completely unexpected, as this personality trait reflects a highly prosocial personality that encompasses traits such as altruism, kindness, trust, and modesty as reflected in the item, "Is considerate and kind to almost everyone" [49,54]. This relationship between agreeableness and leniency was confirmed in both the correlational and the MANCOVA analyses using the severity group as an independent variable.

In the same two analyses, the very same pattern of results was found regarding the relationship between judges' everyday creativity, as measured by the K-DOCS, and their leniency in scoring. Yet, unlike the agreeableness factor, everyday creativity was also found to be a significant and positive predictor of judges' leniency, consistency and discrimination. Similar to agreeableness, everyday creativity is measured by items tapping into highly social, benevolent behaviors, such as, "Helping other people cope with a difficult situation," and "Getting people to feel relaxed and at ease." Yet, it also has an equally represented dimension of intrapersonal creativity [51], accessed by items such as, "Understanding how to make myself happy" and "Maintaining a good balance between my work and my personal life." These aspects of self-awareness, which seem to indicate a capability to positively oversee one's life, appear to be associated with the severity/leniency of scoring, but also (unlike agreeableness) the ability to be consistent and to discriminate different levels of creativity. Thus, although a higher total score for self-assessed creativity was expected to be more relevant to the rating behavior of novice judges, it was not. Instead, our results suggest that it might be useful to take into account the more social characteristics of personality and self-assessed creativity when considering the use of novice judges for creativity. This particular characteristic may play a role in their selection or in the adjustment of their scores. More exploration of this phenomenon is needed.

Before concluding, the current study has some limitations that may be addressed in future extensions of this work. First, it must be noted that our analyses were generally underpowered: the sample size of judges was small and they were asked to rate productions from only one creativity task. In addition, the influence of domain expertise on the raters' evaluations of creativity was confounded by possible gender effects or the effects of the school environment/characteristics. Unfortunately, although variables of SES and ethnicity were collected from each school, numbers in each group proved too small for comparison, making it impossible to explore the homogeneity of the sample across schools.

To improve other weaknesses, more judges within each academic domain would be needed to confirm the results produced here. Importantly, equal numbers of males and females should be recruited within each domain to eliminate the potential confound due to gender differences found in the results, as males

were found to be more severe, less agreeable judges than females, and the judges from computer science were primarily males. In addition, while consistency of scoring is a valuable aspect of rating creativity, more useful information concerning the validity of novice judges' scoring would be gained by comparing the scores generated by this population with the scores of expert judges (once defined).

Finally, it must be acknowledged that the assessment of creative products (resulting from a product-based assessment of creativity) has perhaps limited power to estimate the enduring creativity of the creator. Whereas the creativity of products must reflect some confluence of an individual's knowledge, creativity skills, and motivation—representing a form of potential at a given time in a particular environment [55]—studies on the predictive validity of creative production are sorely needed. Yet, there are very few longitudinal studies of creativity that show the development of creativity across the lifespan, which period may need to be considered, as early inklings of creative potential may or may not manifest as achieved creative outputs (possibly in unpredictable forms) until much later in life [56].

To conclude, the present study has given us some insights as to how a broad conception of creativity, as part of a systems-oriented theory of intelligence, might be successfully assessed in educational settings using novice raters. According to our results, more raters may be needed to reach reliable scores using this method, however, domain differences/areas of interests or developing expertise do not appear to affect creativity ratings. Rater differences instead may be influenced by highly prosocial traits (such as agreeableness) and emotional self-awareness, as reflected in the K-DOCS measure of intrapersonal creativity. While further studies are needed to understand how novice raters might perform when using the classic CAT, this study is meant to propose possible viable methods of assessing creativity that will help push the view of intelligence and intelligence testing outward and the field forward.

Acknowledgments

Many thanks to Vladimir Shpitalnik, who connected the authors with novice judges in the art domain. As always, we extend heartfelt appreciation to Karen Jensen Neff and Charles Neff, whose generous support of the Aurora Project has been extraordinary.

Author Contributions

Mei Tan, Catalina Mourgues, Sascha Hein, and Baptiste Barbot conceived and designed the study. Baptiste Barbot and John MacCormick collected the data. Catalina Mourgues carried out the data analyses, with the support of Sascha Hein and Baptiste Barbot. Mei Tan wrote the paper with input and support from all co-authors, particularly in the Discussion. Elena Grigorenko, as the Primary Investigator, supported all aspects of the work.

Conflicts of interest

The authors declare no conflict of interest.

References

1. Carroll, J.B. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*; Cambridge University Press: New York, NY, USA, 1993.
2. Horn, J.L.; Noll, J. Human cognitive capabilities: Gf-Gc theory. In *Contemporary Intellectual Assessment: Theories, Tests and Issues*; Flanagan, D.P., Genshaft, J.L., Harrison, P.L., Eds.; Guilford Press: New York, NY, USA, 1997; pp. 53–91.
3. McGrew, K.S. The Cattell-Horn Theory of Cognitive Abilities: Past, present, and future. In *Contemporary Intellectual Assessment: Theories, Tests and Issues*; Flanagan, D.P., Harrison, P.L., Eds.; Guilford Press: New York, NY, USA, 2005; pp. 136–181.
4. Siegler, R.S. The other Alfred Binet. *Dev. Psychol.* **1991**, *28*, 179–190.
5. Binet, A.; Simon, T. Application of new methods to the diagnosis of the intellectual level among normal and subnormal children in institutions and in the primary schools. *Annee Psychol.* **1905**, *12*, 245–336.
6. McGrew, K.S. CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* 2**009**, *37*, 1–10.
7. Sternberg, R.J. The theory of successful intelligence. *Rev. Gen. Psychol.* **1999**, *3*, 292–316.
8. Sternberg, R. Applying Psychological Theories to Educational Practice. *Am. Educ. Res. J.* **2008**, *45*, 150–165.
9. Sternberg, R.J. The theory of successful intelligence. *Int. J. Psychol.* **2005**, *39*, 189–202.
10. Sternberg, R.J.; Lubart, T.I.; Kaufman, J.C.; Pretz, J.E. Creativity. In *Cambridge Handbook of Thinking and Reasoning*; Holyoak, K.J., Morrison, R.G., Eds.; Cambridge University Press: Cambridge, UK, 2005; pp. 351–371.
11. Sternberg, R.J.; O' Hara, L.A. Creativity and Intelligence. In *Handbook of Creativity*; Sternberg, R.J., Ed.; Cambridge University Press: New York, NY, USA, 1999; pp. 251–272.
12. Geiser, C.; Mandelman, S.D.; Tan, M.; Grigorenko, E.L. Multi-trait-multimethod assessment of giftedness: An application of the correlated traits—Correlated (methods−1) model. *Struct. Equ. Model. A Multidiscip. J.* **2015**, 1–15, doi:10.1080/10705511.2014.937792.
13. Kornilov, S.A.; Tan, M.; Elliott, J.G.; Sternberg, R.J.; Grigorenko, E.L. Gifted identification with Aurora: Widening the spotlight. *J. Psychoeduc. Assess.* **2011**, doi:10.1177/0734282911428199.
14. Tan, M.; Mourgues, C.; Bolden, D.S.; Grigorenko, E.L. Making Numbers Come to Life: Two Scoring Methods for Creativity in Aurora's Cartoon Numbers. *J. Creat. Behav.* **2014**, *48*, 25–43.
15. Amabile, T.M. Social psychology of creativity: A consensual assessment technique. *J. Personal. Soc. Psychol.* **1982**, *43*, 997–1013.
16. Hennessey, B.A.; Amabile, T.M. Consensual assessment. In *Encyclopedia of Creativity*; Runco, M.A., Pritzker, S.R., Eds.; Academic Press: San Diego, CA, USA, 1999; pp. 347–359.
17. Myford, C.M.; Wolfe, E.W. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *J. Appl. Meas.* **2003**, *4*, 386–422.
18. Cronbach, L. Processes affecting scores on "understanding of others" and "assumed similarity". *Psychol. Bull.* **1955**, *52*, 177–193.
19. Kaufman, J.C.; Baer, J.; Cole, J.C.; Sexton, J.D. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creat. Res. J.* **2008**, *20*, 171–178.

20. Silvia, P.J.; Winterstein, B.P.; Willse, J.T.; Barona, C.M.; Cram, J.T.; Hess, K.I.; Martinez, J.L.; Richard, C.A. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychol. Aesthet. Creat. Arts* **2008**, *2*, 68–85.
21. Storme, M.; Lubart, T. Conceptions of creativity and relations with judges' intelligence and personality. *J. Creat. Behav.* **2012**, *46*, 138–149.
22. Runco, M.A.; Jaeger, G.J. The Standard Definition of Creativity. *Creat. Res. J.* **2012**, *24*, 92–96.
23. Lubart, T.; Sternberg, R.J. Creativity. In *Thinking and Problem Solving*; Sternberg, R.J., Ed.; Academic Press: San Diego, CA, USA, 1994.
24. Sternberg, R.J. The Nature of Creativity. *Creat. Res. J.* **2006**, *18*, 87–98.
25. Hennessey, B.A.; Amabile, T.M. Creativity. *Annu. Rev. Psychol.* **2010**, *61*, 569–598.
26. Barbot, B.; Besançon, M.; Lubart, T. Creative potential in educational settings: its nature, measure, and nurture. *Int. J. Prim. Elem. Early Year Educ.* **2015**, *43*, 371–381.
27. Barbot, B.; Tinio, P.P. Where is the "g" in creativity? A specialization–differentiation hypothesis. *Front. Hum. Neurosci.* **2014**, *8*, 1041, doi:10.3389/fnhum.2014.01041.
28. Runco, M.A. The creativity of children's art. *Child Study J.* **1989**, *19*, 177–189.
29. Kaufman, J.C.; Lee, J.; Baer, J.; Lee, S. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Think. Skills Creat.* **2007**, *2*, 96–106.
30. Baer, J.; Kaufman, J.C.; Gentile, C.A. Extension of the consensual assessment technique to nonparallel creative products. *Creat. Res. J.* **2004**, *16*, 113–117.
31. Kaufman, J.C.; Baer, J. Beyond new and appropriate: Who decides what is creative? *Creat. Res. J.* **2012**, *24*, 83–91.
32. Hennessey, B.A. The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creat. Res. J.* **1994**, *7*, 193–208.
33. Cropley, D.H.; Kaufman, J.C. Measuring Functional Creativity: Non-Expert Raters and the Creative Solution Diagnosis Scale. *J. Creat. Behav.* **2012**, *46*, 119–137.
34. Haller, C.S.; Courvoisier, D.S.; Cropley, D.H. Perhaps there is accounting for taste: Evaluating the creativity of products. *Creat. Res. J.* **2011**, *23*, 99–109.
35. White, A.; Shen, F.; Smith, B.L. Judging advertising creativity using the creative product semantic scale. *J. Creat. Behav.***2002**, *36*, 241–253.
36. Baer, J.; McKool, S. Assessing creativity using the consensual assessment. In *Handbook of Assessment Technologies, Methods and Applications in Higher Education*; Information Science Reference: Hershey, PA, USA, 2009; pp. 65–77.
37. Storme, M.; Myszkowski, N.; Çelik, P.; Lubart, T. Learning to judge creativity: The underlying mechanisms in creativity training for non−expert judges. *Learn. Individ. Differ.* **2014**, *32*, 19–25.
38. Runco, M.A.; Charles, R.E. Judgments of originality and appropriateness as predictors of creativity. *Personal. Individ. Differ.* **1993**, *15*, 537–546.
39. Caroff, X.; Besançon, M. Variability of creativity judgments. *Learn. Individ. Differ.* **2008**, *18*, 367–371.
40. Priest, T. The reliability of three groups of judges' assessments of creativity under three conditions. *Bull. Counc. Res. Music Educ.* **2006**, *Winter 2006*, 47–60.
41. Kaufman, J.C.; Baer, J.; Cole, J.C. Expertise, domains, and the consensual assessment technique. *J. Creat. Behav.* **2009**, *43*, 223–233.

42. Plucker, J.A.; Kaufman, J.C.; Temple, J.S.; Qian, M. Do experts and novices evaluate movies the same way? *Psychol. Mark.* **2009**, *26*, 470–478.
43. Chen, C.; Kasof, J.; Himsel, A.J.; Greenberger, E.; Dong, Q.; Xue, G. Creativity in Drawings of Geometric Shapes A Cross-Cultural Examination with the Consensual Assessment Technique. *J. Cross Cult. Psychol.* **2002**, *33*, 171–187.
44. Niu, W.; Sternberg, R.J. Cultural influences on artistic creativity and its evaluation. *Int. J. Psychol.* **2001**, *36*, 225–241.
45. Hood, R.W., Jr. Rater Originality and the Interpersonal Assessment of Levels of Originality. *Sociometry* **1973**, *36*, 80–88.
46. Pearson, P.H. Relationships between global and specified measures of novelty seeking. *J. Consult. Clin. Psychol.* **1970**, *34*, 199–204.
47. Piedmont, R.L.; McRae, R.R.; Costa, P.T. An assessment of the Edwards personal preference schedule from the perspective of the five factor model. *J. Personal. Assess.* **1992**, *58*, 67–78.
48. Kozbelt, A.; Durmysheva, Y. Understanding Creativity Judgments of Invented Alien Creatures: The Roles of Invariants and Other Predictors. *J. Creat. Behav.* **2007**, *41*, 223–248.
49. John, O.P.; Srivastava, S. The Big Five trait taxonomy: Hstory, measurement, and theoretical perspective. In *Handbook of Personality: Theory and Research*; Pervin, L.A., John, O.P., Eds.; Guilford Press: New York, NY, USA, 1999; pp. 102–138.
50. John, O.P.; Naumann, L.P.; Soto, C.J. Pardigm shift to the integrative Big-Five Trait Taxonomy: HIstory, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research*; John, O.P., Robns, R.W., Pervin, L.A., Eds.; Guilford Press: New York, NY, USA, 2008; pp. 114–158.
51. Kaufman, J.C. Counting the muses: Development of the Kaufman Domains of Creativity Scale (K−DOCS). *Psychol. Aesthet. Creat. Arts* **2012**, *6*, 298–308.
52. Linacre, J.M. *Facets: Rasch Measurement Computer Program (version 3.65.0)*; Winsteps: Chicago, IL, USA, 2009.
53. Kaufman, J.C.; Baer, J.; Cropley, D.H.; Reiter-Palmon, R.; Sinnett, S. Furious activity *vs.* understanding: How much expertise is needed to evaluate creative work? *Psychol. Aesthet. Creat. Arts* **2013**, *7*, 332–340.
54. Costa, P.T.; McCrae, R.R. Four ways five factors are basic. *Personal. Individ. Differ.* **1992**, *13*, 653–665.
55. Hennessey, B.A. Comment on "The Psychology of creativity: A critical reading" by Vlad Petre Glăveanu. *Creat. Theor. Res. Appl.* **2015**, *2*, 32–37.
56. Lubart, T.; Zenasni, F.; Barbot, B. Creative potential and its measurement. *Int. J. Talent Dev. Creat.* **2013**, *1*, 41–51.