A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking.

By: Paul J. Silvia, Christopher Martin, & Emily C. Nusbaum

Paul J. Silvia, Christopher Martin, & Emily C. Nusbaum (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4(2)*, 79–85.

Abstract:

Creativity assessment commonly uses open-ended divergent thinking tasks. The typical methods for scoring these tasks (uniqueness scoring and subjective ratings) are time-intensive, however, so it is impractical for researchers to include divergent thinking as an ancillary construct. The present research evaluated snapshot scoring of divergent thinking tasks, in which the set of responses receives a single holistic rating. We compared snapshot scoring to top-two scoring, a time-intensive, detailed scoring method. A sample of college students (n = 226) completed divergent thinking tasks and measures of personality and art expertise. Top-two scoring had larger effect sizes, but snapshot scoring performed well overall. Snapshot scoring thus appears promising as a quick and simple approach to assessing creativity.

Keywords: creativity | divergent thinking | validity | psychological assessment | openness to experience | psychology

Article:

Like parents, ombudsmen, and city council members, researchers are used to compromise. All assessment involves trade-offs between a method's evidence for validity and its cost. Many of the best assessment tools are costly in terms of administration time, expertise, technology, personnel-hours, and infrastructure. For this reason, many constructs have a range of available tools. A person's typical mood can be assessed with week-long experience-sampling methods or with brief self-report scales. Clinical symptoms can be assessed with face-to-face clinical interviews or with brief self-report screening scales. Even within a method, researchers can usually find a range of options. Personality researchers, for example, could choose the 300-item NEO-PI (Costa & McCrae, 1992), the 60-item FFI (Costa & McCrae, 1992), a 20-item IPIP scale (Donnellan, Oswald, Baird, & Lucas, 2006), or even one of two 10-item scales (Gosling et al., 2003 and Rammstedt and John, 2007).

The present research appraises a quick and simple method for assessing individual differences in creativity. Creativity research typically uses divergent thinking tasks to measure variation in creative abilities and potential (Kaufman et al., 2008, Plucker and Renzulli, 1999 and Runco, 2007), but the traditional methods of coding and scoring these tasks are costly in terms of time and personnel. As a result, creativity is hard to include as a secondary or exploratory construct in a research project. We compare this brief, simple method—known as snapshot scoring—to a more time-consuming, detailed method (Silvia et al., 2008). If the brief method performs well, it may be a useful tool for researchers interested in assessing creativity.

1. Major approaches to scoring divergent thinking

Divergent thinking tasks assess creativity by asking people to generate ideas, which are then scored to capture variability in creativity (Plucker & Renzulli, 1999). For example, unusual uses tasks ask people to generate unusual uses for common objects, such as bricks, knives, boxes, shoes, and paper clips. Over the decades, researchers have developed many methods for deriving scores from these tasks (Kaufman et al., 2008). Two scoring methods—uniqueness scoring and subjective ratings—have been used the most extensively.

1.1. Uniqueness scoring

Uniqueness scoring, formalized by Wallach and Kogan (1965) in their landmark research, is straightforward. Researchers compile all of the responses to a task and then assign each response a 0 or a 1. If a response is literally unique—if only one person in the sample gave the response—then it gets a 1. All other responses get 0s, regardless of how often they appeared. An appeal of this method is its implied definition of creativity: creative ideas are unique ideas. In this system, researchers focus on two scores: uniqueness (the number of unique responses) and fluency (the total number of responses), which indicate the quality and quantity of divergent thinking.

All research is at least slightly painful, but uniqueness scoring is uniquely painful. The responses must be transcribed into a lexicon, which may have tens of thousands of responses. Each response must then be compared with each other response to see if it appears only once or not. Along the way, many judgment calls must be made, such as whether “a brick walkway” and “a brick path” are unique responses and whether incomprehensible yet unique responses (e.g., “You know, that thing that hangs on a tree trunk. You know what I’m talking about, right?”) deserve a point.

Uniqueness scoring has two major problems that have motivated the development of alternative assessment systems. The first problem is that uniqueness scores are highly correlated with fluency scores. Researchers pointed this out soon after Wallach and Kogan (1965) published their landmark tasks and scoring methods (e.g., Clark & Mirels, 1970), and later research further demonstrated that uniqueness and fluency are essentially confounded (Hocevar, 1979a, Hocevar, 1979b and Hocevar and Michael, 1979). This confounding is apparent even in the field's "gold standard" samples. In a reanalysis of Wallach and Kogan's data, Silvia (2008b) found a correlation of $\beta = .89$ between the latent fluency and uniqueness variables. In the large norm sample for the Torrance Tests of Creative Thinking (Torrance, 2008), fluency and uniqueness correlated equally highly (median $r = .88$). When two variables covary this strongly, it is hard to believe that they convey distinct information. Many contemporary researchers, in fact, use only the fluency scores as markers of creativity (e.g., Batey et al., 2009 and Preckel et al., 2006).

The second problem is that uniqueness scores are strongly biased by sample size—as the sample becomes larger, creative responses become less common. Creativity thus becomes harder to detect, and the sample's estimated level of creativity declines, as the sample increases. In a small sample of 20 people, for example, most responses will be unique because the pool of responses is small. If 200 cases are added, however, many formerly unique responses will no longer be unique in the larger pool. In a huge sample, very few responses will be unique; it is theoretically possible, in fact, for no responses to be unique. Uniqueness scoring thus introduces a powerful sample dependence in its estimates of creativity, and this is bad for obvious reasons. Any assessment method that fares poorly in large samples is undesirable for basic research, unsuitable for large-scale testing programs, and unacceptable for high-stakes purposes (e.g., placement into gifted-education programs).

1.2. Subjective ratings of responses

A second assessment tradition uses judges to provide subjective ratings of creativity. The best-known example is the consensual assessment technique (CAT; Amabile, 1982), which is used to appraise creative products. In a typical CAT study, people are asked to generate poems, collages, or stories; experts in the area then judge the products on various dimensions (e.g., creativity, technical skill) according to their personal definition of creativity (Kaufman et al., 2008). Recent work has loosened the classic CAT to include diverse kinds of products, raters, and judgments (Baer et al., 2004, Kaufman et al., 2005 and Kaufman et al., 2007).

In divergent thinking research, it is difficult to conceive of what an “expert” divergent thinking judge would be. Instead of noted experts, stoic research assistants are usually pressed into service for the tedious process of rating divergent thinking responses. The responses are usually rated on a creative-quality scale, such as a 5-point “not at all creative” to “very creative” response format (Silvia et al., 2008). Unlike the CAT, which avoids training the raters, divergent thinking research usually provides some training, guidelines, or rubrics for judges, thus increasing between-judge agreement.

Subjective scoring of divergent thinking tasks dates at least to Guilford's seminal research; many of his studies used subjective ratings of originality, remoteness, and cleverness (e.g., Christensen et al., 1957 and Wilson et al., 1953). Since then, many studies have used some kind of subjective quality rating. The most common approach is to have raters give a score to each response (e.g., Gilhooly et al., 2007, Harrington, 1975 and Silvia and Phillips, 2004). These scores can then be averaged for an overall score for the task.

When scoring each response, researchers can score other features of the responses and test if those features covary with the creativity ratings. For example, people's later responses are usually more creative than their first few responses (Christensen et al., 1957), responses that people generated on the spot are more creative than responses retrieved from memory (Gilhooly et al., 2007), and the two responses that people picked as their “top two” are more creative than the rest (Silvia, 2008c). Responses can also be coded for their length, elaborateness, concreteness, cleverness, and their clustering into classes—these open-ended tasks provide a lot of information.

2. Snapshot scoring: holistic judgments of creativity

Subjective ratings overcome uniqueness scoring's two problems: they are not confounded with fluency, and they are not biased by large samples.1 They do share, however, the painfulness of uniqueness scoring. The responses must be entered into a database, coded for internal features (e.g., a response's serial position, whether a response was picked as one of the best responses), and then scored response by response. A large-sample study with several tasks and several raters can have around 50,000 ratings. Software tools—such as Pretz and Link's (2008) clever Creative Task Creator—can simplify the process, but based on our experience, researchers who wish to use subjective scoring will need at least three committed raters, several weeks for the ratings, thousands of milligrams of caffeine, and either an Atari 2600 or a Nintendo Wii for the lab.

A simple, quick method for scoring divergent thinking tasks would thus be helpful. For many studies, creativity is secondary to the project, so it is inefficient to devote hundreds of personnel-hours to coding and scoring a peripheral construct. Some researchers may lack the time or research infrastructure to commit to including creativity tasks in their research. One promising method that has appeared a few times in past research is snapshot scoring of divergent thinking tasks. In this method, raters view the entire set of responses and give a single holistic rating to the set. No internal coding or scoring is performed. For pencil-and-paper tasks, the responses can be scored from the handwritten sheets, thus avoiding the transcription needed for response-by-response scoring.

A few studies have explored the value of holistic scoring of divergent thinking tasks. In one study (Runco & Mraz, 1992), 24 adolescents completed divergent thinking tasks, and 20 college students rated the creativity of each set of responses. The raters used a 1–7 scale, with the constraint that the ratings had to approximate a normal distribution. The holistic scores had good reliability, but the study did not collect evidence for the scores' validity. Moreover, the study was clearly preliminary, given the small sample (n = 24) and the unusually large number of raters to participants.

In a later study, Mouchiroud and Lubart (2001) asked adults to give overall scores to divergent thinking sets provided by a sample of 70 children. One group of three adults scored the tasks for "creativity"; another group of three adults scored the tasks for "originality." The raters used a 1–7 scale. These scores had good internal consistency, and they covaried highly with each other (r = .76). No evidence for concurrent validity was collected, however.

3. The present study

Snapshot scoring is certainly quick and simple, but does it work? Past research has explored snapshot scoring (Mouchiroud and Lubart, 2001 and Runco and Mraz, 1992), but the evidence to date is clearly preliminary. There is some evidence for score reliability, but thus far there is little evidence for the scores' validity. The present research evaluated the merit of snapshot scoring in more detail. Do the scores display acceptable levels of reliability and validity? Instead of appraising snapshot scoring in isolation, we compared it to an established and time-consuming method—top-two scoring, which transcribes and scores all responses but uses only people's best two responses to assess their level of creativity (Silvia et al., 2008). Directly comparing the evidence for methods' validity allows researchers to estimate the likely consequences of using the quick snapshot scoring instead of the complicated top-two scoring.

The data were taken from a large-scale project that explored creativity's relationships with personality, intelligence, interests, and creative accomplishments over time (Silvia, 2007, Silvia, 2008b and Silvia et al., 2008). There is thus a broad range of constructs that can be used to appraise the concurrent validity of snapshot scores. In the present study, we evaluated the relationships between the two scoring methods and constructs related to personality (the Big Five domains) and expertise in the arts. By rescoring existing data, we hope to foster a cumulative and comparative approach to measuring creativity. Researchers interested in developing new scoring methods can obtain the raw data and responses and thus compare their method to both snapshot scoring and top-two scoring.

4. Method

4.1. Participants

The original data set was collected as part of the Creativity and Cognition project. It consists of data from 226 students—178 women, 48 men—enrolled in General Psychology at the University of North Carolina at Greensboro. More details about the sample are available in the original article (Silvia et al., 2008).

4.2. Procedure

The people who participated in the study completed a wide range of tasks and self-report scales.

4.2.1. Divergent thinking

We used two unusual uses tasks—uses for a brick and for a knife—to measure creativity. Unusual uses tasks seem to perform better than other kinds of divergent thinking tasks (Silvia et al., 2008, Study 1). People were told that the tasks concerned creative thinking and that they should try to come up with creative responses. Instructing people to be creative expands the between-person variance in creativity and increases the validity of the scores (Harrington, 1975). They had 3 min for each task. After the task, the experimenter asked people to read their responses and to circle the two that they thought were their best responses. By asking people to be creative and allowing them to indicate their most creative responses, this method approximates “maximal assessment” of creativity, in which people's best level (rather than typical level) is assessed (Runco, 1986).

All of the responses were transcribed into a database; spelling errors were silently corrected. Three research assistants rated each response. The responses were sorted alphabetically within each task, so the raters were unaware of whether a response was picked as a top-two response, the response's position in the set, the number of responses in the set, the person's other responses on the task, or any information whatsoever about the person's responses to the questionnaires and other tasks. Each response was rated on a 1–5 scale that ranged from “not at all creative” to “very creative.” Guilford's notions of creative ideas as uncommon, remote, and clever (Wilson et al., 1953) were used as scoring guidelines; the exact descriptions are published in an appendix to Silvia et al. (2008). The two tasks yielded over 3224 responses, so the group of raters provided 9672 ratings.

4.2.2. Snapshot scoring

Three new raters conducted the snapshot scoring. These raters had not conducted the top-two ratings for this sample, but they had conducted top-two scoring for a different data set. After reviewing the scoring guidelines, the raters read and scored the divergent thinking tasks directly from the original response sheets. The raters were thus aware of peripheral information (e.g., handwriting, any spelling errors, whether the participant used a sparkly lavender gel pen) as well as task-relevant information (e.g., number and serial order of responses, which responses were circled as top-two responses), but they were unaware of all responses to the other questionnaires and tasks. Each sheet—each set of responses—was given a single number. The raters used a 1–5 scale that ranged from “not at all creative” to “very creative.” The group of raters provided 1356 scores, around 14% of the amount required for the top-two scoring method.

4.2.3. Personality

We measured the Big Five personality domains with 3 scales: the 60-item Five Factor Inventory (Costa & McCrae, 1992), a 50-item International Personality Item Pool scale (Goldberg et al., 2006), and a 10-item brief scale (Gosling et al., 2003). Each item used a 5-point response format.

4.2.4. Art expertise

We assessed people's expertise in the arts two ways. First, we used Smith and Smith's (2006) aesthetic fluency scale, a 10-item self-report scale that measures people's knowledge of figures and terms from art history. People indicate how familiar they are with the person or term on a 5-point scale. To expand the scale's content, we added five items related to creative writing and 5 items related to the decorative arts. Although new, this scale has performed well in several studies (Silvia, 2007, Silvia and Barona, 2009 and Smith and Smith, 2006).

Second, we classified each student's major as a creative college major (scored as 1) or a conventional college major (scored as 0). About 9% of the sample had a creative major (e.g., fine arts, interior architecture, music performance). Although coarse, this scoring broadly captures whether students have creative goals, career aspirations, and occupational interests (see Silvia et al., 2008, for details).

5. Results

5.1. Model specification and analytic method

We used Mplus 5.2 for all analyses. The models were estimated with full-information maximum likelihood with robust standard errors (MLR). Few observations were missing; most analyses had no missing data, and the matrix cells with the most missing data were nevertheless 99.1% complete.

For snapshot scores, we specified creativity as a higher order latent variable defined by lower order brick and knife variables. The variance of creativity was fixed to 1, and the paths to brick and knife were constrained to be equal for identification. For the brick and knife variables, the paths to the three raters' scores were the indicators; the brick and knife variances were fixed to 1. The raters' scores were modeled as ordered-categorical (ordinal) variables: they were highly skewed, and some response options (4 and 5) were rarely chosen by the raters. (Readers familiar with item-response theory will recognize this as a graded response model for polytomous outcomes.) Modeling the snapshot ratings as ordinal only trivially changed the effects, but it did restrict the kinds of model fit statistics that could be reported (see Skrondal & Rabe-Hesketh, 2004, for details on generalized CFA models).

The faceted structure of the snapshot scoring—three raters crossed with two tasks—and the ordinal nature of the ratings make conventional reliability indices (e.g., Cronbach's alpha) inappropriate. Nevertheless, an analogous statistic for latent variable models is maximal reliability (Drewes, 2000), which represents “the degree to which the indicators can capture information about the underlying factor” (Gagné & Hancock, 2006, p. 68). The higher order creativity variable ($H = .83$) and the lower order brick ($H = .84$) and knife ($H = .88$) variables displayed good reliability.

For top-two scores, we used the same model specification that we used in past work (Silvia, 2008a and Silvia et al., 2008). Creativity was a higher order latent variable indicated by lower order brick and knife variables. The variance of creativity was fixed to 1, and the paths were constrained to be equal for identification. For brick and knife, the three raters' scores were the indicators. The paths to Rater 2 were fixed to 1, and the paths for Raters 1 and 3 were constrained to be equal because they were very similar. This model fit well: CFI = .990, RMSEA = .034, SRMR = .050.

5.2. Personality and creativity

For the personality model, we estimated the effects of the Big Five variables on snapshot scores and top-two scores. The Big Five variables were modeled as latent variables; the three scales were the indicators, and the path to the IPIP scale was fixed to 1. Table 1 depicts the standardized effects, which can be interpreted as effect sizes, and their confidence intervals; Table 2 depicts the variance explained. Note that the top-two effects appeared in Silvia et al. (2008, Study 2).[2]

Table 1. Summary of effects for snapshot scores and top-two scores.

Model	Snapshot scores		Top-two scores	
	β	CI	β	CI
Personality model				
Neuroticism	−.041	−.220 to .138	.026	−.212 to .264
Extraversion	−.153	−.333 to .028	−.054	−.361 to .253
Openness to experience	.330	.131 to .528	.578	.245 to .910
Agreeableness	.081	−.096 to .258	.242	−.013 to .497
Conscientiousness	−.288	−.469 to −.108	−.461	−.736 to −.186
Art expertise model				
Aesthetic fluency	.090	−.080 to .261	.132	−.132 to .396
Creative college major	.974	.538 to 1.409	1.664	.828 to 2.501

Note: n = 226. The coefficients are standardized regression weights, which represent effect sizes. CI = 95% symmetric confidence intervals around β. The coefficient for Creative College Major, a binary predictor, is *Y*-standardized: it is the standard deviation change in the outcome when the predictor shifts from 0 to 1 (Long, 1997).

Table 2. Summary of variance explained in creativity scores.

	Snapshot scores	Top-two scores
Personality model	15.5%	48.3%
Art expertise model	8.8%	26.6%

The snapshot scores performed well: they were significantly predicted by openness to experience and conscientiousness, two variables that consistently appear in research. As in past research, openness and conscientiousness had opposite effects on creativity (Batey and Furnham, 2006 and Feist, 1998). Nevertheless, the effects for the top-two scores were larger. Overall, the Big Five explained a lot of variance in snapshot scores (15.5%) but much more in top-two scores (48.3%).

5.3. Art expertise and creativity

Our next model examined art expertise, assessed by aesthetic fluency scores and people's college majors. We modeled aesthetic fluency as a latent variable indicated by the three subscales (fine arts, creative writing, and decorative arts); the path to the fine arts variable was fixed to 1. Table 1 depicts the effects. Aesthetic fluency had only a small effect on both snapshot scores and top-two scores, although the effect size was larger for top-two scores.

Having a creative college major, in contrast, had a large effect on creativity. Because the predictor is binary, the coefficient is Y-standardized (Long, 1997): it shows the standard deviation change in the outcome when the predictor shifts from 0 (conventional major) to 1 (creative major). For snapshot scoring, the coefficient of .97 means that people with conventional and creative majors differed in creativity by about 1 standard deviation. For top-two scoring, the effect was much larger: people with conventional and creative majors differed in creativity by about 1.7 standard deviations.

6. General discussion

Creativity assessment is time-consuming. Researchers who study creativity cannot complain—it is what they signed up for, and many constructs are harder to assess than creativity. The problem, however, is that researchers interested in other topics are discouraged from including creativity as an ancillary, secondary, or exploratory construct. The commitment of time and personnel is a big barrier to exploring creativity's relationships with a wide network of other constructs; this ultimately stunts the growth of knowledge in the field. It would thus be nice to have a simpler and faster method of creativity scoring, one that strikes an appropriate balance between convenience and psychometric effectiveness.

The present research explored snapshot scoring—giving a holistic rating to a set of divergent thinking responses—as a quick and simple method. Overall, this method appears promising. First, the factor structure of the snapshot scores showed good construct reliability (Drewes, 2000)—all of the H values were over .80, so good score reliability was achieved with only three raters and two tasks. Second, good evidence for concurrent validity was found. Snapshot scores covaried positively with openness to experience and people's college majors; they covaried negatively with conscientiousness, thus replicating much past research (Batey and Furnham, 2006 and Feist, 1998).

At the same time, top-two scoring—the more complex, costly, and time-consuming method—generally performed better than the quick and simple snapshot method. More variance was accounted for in top-two scores than in snapshot scores. It is reassuring, in a sense, that the tedious yet detailed method generally performed better—this shows that there is merit in taking the extra time and effort to score the tasks more precisely. Furthermore, snapshot scoring does not afford testing certain kinds of hypotheses, such as hypotheses about the order or clustering of responses. Researchers interested in internal features of the responses must score each response individually.

By comparing the two scoring methods, we can illuminate the nature of the compromise between convenience and psychometric effectiveness. For example, openness to experience predicted both snapshot scores and top-two scores, but it predicted top-two scores much more strongly. For studies of personality, it seems that the extra effort is justified. For art expertise, in contrast, researchers may believe that gain in effect size is not worth the larger cost in time and effort. Such decisions are subjective, but they can be informed by comparative studies of validity. We encourage future researchers to score their data according to several methods—or to rescore the responses used in this research—thus fostering a cumulative literature on the effectiveness of creativity assessment.

Acknowledgements

We thank Chris Barona, Josh Cram, Karl Hess, Jenna Martinez, and Crystal Richard for their help in collecting the original data. This work was supported by an Undergraduate Research Assistantship to Christopher Martin.

References

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

J. Baer, J.C. Kaufman, C.A. Gentile. Extension of the consensual assessment technique to nonparallel creative products. Creativity Research Journal, 16 (2004), pp. 113–117

M. Batey, T. Chamorro-Premuzic, A. Furnham. Intelligence and personality as predictors of divergent thinking: The role of general, fluid and crystallised intelligence. Thinking Skills and Creativity, 4 (2009), pp. 60–69

M. Batey, A. Furnham. Creativity, intelligence, and personality: A critical review of the scattered literature. Genetic, Social, and General Psychology Monographs, 132 (2006), pp. 355–429

P.R. Christensen, J.P. Guilford, R.C. Wilson. Relations of creative responses to working time and instructions. Journal of Experimental Psychology, 53 (1957), pp. 82–88

P.M. Clark, H.L. Mirels. Fluency as a pervasive element in the measurement of creativity. Journal of Educational Measurement, 7 (1970), pp. 83–86

P.T. Costa Jr., R.R. McCrae. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. PAR, Odessa, FL (1992)

M.B. Donnellan, F.L. Oswald, B.M. Baird, R.E. Lucas. The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. Psychological Assessment, 18 (2006), pp. 192–203

D.W. Drewes. Beyond the Spearman–Brown: A structural approach to maximal reliability

Psychological Methods, 5 (2000), pp. 214–227

G.J. Feist. A meta-analysis of personality in scientific and artistic creativity. Personality and Social Psychology Review, 2 (1998), pp. 290–309

P. Gagné, G.R. Hancock. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behavioral Research, 41 (2006), pp. 65–83

K.J. Gilhooly, E. Fioratou, S.H. Anthony, V. Wynn. Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. British Journal of Psychology, 98 (2007), pp. 611–625

L.R. Goldberg, J.A. Johnson, H.W. Eber, R. Hogan, M.C. Ashton, C.R. Cloninger et al. The international personality item pool and the future of public-domain personality assessment. Journal of Research in Personality, 40 (2006), pp. 84–96

S.D. Gosling, P.J. Rentfrow, W.B. Swann Jr. A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37 (2003), pp. 504–528

D.M. Harrington. Effects of explicit instructions to “be creative” on the psychological meaning of divergent thinking test scores. Journal of Personality, 43 (1975), pp. 434–454

D. Hocevar. A comparison of statistical infrequency and subjective judgment as criteria in the measurement of originality. Journal of Personality Assessment, 43 (1979), pp. 297–299

D. Hocevar. Ideational fluency as a confounding factor in the measurement of originality. Journal of Educational Psychology, 71 (1979), pp. 191–196

D. Hocevar, W.B. Michael. The effects of scoring formulas on the discriminant validity of tests of divergent thinking. Educational and Psychological Measurement, 39 (1979), pp. 917–921

J.C. Kaufman, C.A. Gentile, J. Baer. Do gifted student writers and creative writing experts rate creativity the same way? Gifted Child Quarterly, 49 (2005), pp. 260–265

J.C. Kaufman, J. Lee, J. Baer, S. Lee. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. Thinking Skills and Creativity, 2 (2007), pp. 96–106

J.C. Kaufman, J.A. Plucker, J. Baer Essentials of creativity assessment. Wiley, Hoboken, NJ (2008)

J.S. Long. Regression models for categorical and limited dependent variables. Sage, Thousand Oaks, CA (1997)

C. Mouchiroud, T. Lubart. Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. Journal of Genetic Psychology, 162 (2001), pp. 382–401

J.A. Plucker, J.S. Renzulli. Psychometric approaches to the study of human creativity. R.J. Sternberg (Ed.), Handbook of creativity, Cambridge University Press, New York (1999), pp. 35–61

F. Preckel, H. Holling, M. Wiese. Relationship of intelligence and creativity in gifted and non-gifted students: An investigation of threshold theory. Personality and Individual Differences, 40 (2006), pp. 159–170

J.E. Pretz, J.A. Link. The creative task creator: A tool for the generation of customized, Web-based creativity tasks. Behavior Research Methods, 40 (2008), pp. 1129–1133

B. Rammstedt, O.P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, 41 (2007), pp. 203–212

M.A. Runco. Maximal performance on divergent thinking tests by gifted, talented, and nongifted students. Psychology in the Schools, 23 (1986), pp. 308–315

M.A. Runco. Creativity. Elsevier, Amsterdam (2007)

M.A. Runco, W. Mraz. Scoring divergent thinking tests using total ideational output and a creativity index. Educational and Psychological Measurement, 52 (1992), pp. 213–221

P.J. Silvia. Knowledge-based assessment of expertise in the arts: Exploring aesthetic fluency. Psychology of Aesthetics, Creativity, and the Arts, 1 (2007), pp. 247–249

P.J. Silvia. Another look at creativity and intelligence: Exploring higher-order models and probable confounds. Personality and Individual Differences, 44 (2008), pp. 1012–1021

P.J. Silvia. Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). Creativity Research Journal, 20 (2008), pp. 34–39

P.J. Silvia. Discernment and creativity: How well can people identify their most creative ideas? Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 139–146

P.J. Silvia, C.M. Barona. Do people prefer curved objects? Angularity, expertise, and aesthetic preference. Empirical Studies of the Arts, 27 (2009), pp. 25–42

P.J. Silvia, A.G. Phillips. Self-awareness, self-evaluation, and creativity. Personality and Social Psychology Bulletin, 30 (2004), pp. 1009–1017

P.J. Silvia, B.P. Winterstein, J.T. Willse, C.M. Barona, J.T. Cram, K.I. Hess et al. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 68–85

A. Skrondal, S. Rabe-Hesketh. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Chapman & Hall/CRC, Boca Raton, FL (2004)

L.F. Smith, J.K. Smith. The nature and growth of aesthetic fluency. P. Locher, C. Martindale, L. Dorfman (Eds.), New directions in aesthetics, creativity, and the arts, Baywood, Amityville, NY (2006), pp. 47–58

E.P. Torrance. Torrance tests of creative thinking: Norms-technical manual, verbal forms A and B. Scholastic Testing Service, Bensenville, IL (2008)

M.A. Wallach, N. Kogan. Modes of thinking in young children: A study of the creativity–intelligence distinction.Holt, Rinehart, & Winston, New York (1965)

R.C. Wilson, J.P. Guilford, P.R. Christensen. The measurement of individual differences in originality. Psychological Bulletin, 50 (1953), pp. 362–370

Corresponding author at: Department of Psychology, P.O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC 27402-6170, United States. Tel.: +1 336 256 0007; fax: +1 336 334 5066.

[1] As an aside, subjective ratings of creativity could in principle meet the Rasch ideal of "specific objectivity," in which trait estimates are independent of the tasks and items used to measure the trait. Approximating scores with specific objectivity would probably require a multidimensional item response model that separated the influence of raters from the influence of the latent creativity trait. Because of their unusual sample dependence, uniqueness scores cannot in principle achieve specific objectivity.

[2] Fastidious readers will notice some small differences in the effects reported here and the effects reported earlier. For the personality models, a few top-two effects differ trivially because the Silvia et al. (2008) analyses used simple maximum likelihood (ML) estimation.