Psychology of Aesthetics, Creativity, and the Arts
2008, Vol. 2, No. 2, 89–92

1931-3896/08/$12.00 DOI: 10.1037/1931-3896.2.2.89

Commentary: Divergent Thinking Tests Have Problems, But This Is Not the Solution

John Baer
Rider University

It is certainly true, as Silvia et al. (2008) write, that "after half a century of research, the evidence for global creative ability ought to be better" (p. 68). The authors believe—incorrectly, I think—that the reason that divergent thinking tests have not done a better job can be found in the various scoring systems that have been used when assessing divergent thinking ability. I have presented evidence elsewhere that creativity is *not* a general ability or set of traits or dispositions that can be applied across domains (Baer, 1991, 1993, 1994a, 1994b, 1998). In those studies, I used Amabile's (1982, 1996) Consensual Assessment Technique (which is the basis for the subjective scoring technique proposed by Silvia et al. [2008]) to judge the creativity of a wide range of artifacts. What I found was that there is little correlation among the creativity ratings received by subjects across domains, and what little there is tends to disappear if an IQ test is also given and variance attributable to intelligence is first removed.

If creativity is not a generic, all-purpose kind of skill—that is, if whatever it is that leads to creativity in, say, writing poetry does not also enhance creativity in teaching, creativity in cooking, and creativity in any other kind of activity—then we should not be surprised to find that tests of general creativity ability lack validity. In the arena of IQ testing, it has been shown that whatever it is that IQ tests measure is positively correlated with actual performance in a wide, domain-transcending range of tasks, but this is not the case for divergent-thinking testing. Perhaps the reason that "after half a century of research, the evidence for global creative ability ought to be better" (p. 68) is that, unlike intelligence, there simply is no general creativity skill to be measured. It may be that creativity is largely domain specific, a conclusion for which the evidence from the assessment of actual creative products is quite convincing. If so, then the construct of general creativity is a false one, and it doesn't matter how you score creativity tests. Such tests can never be valid for the simple reason that they purport to measure something that doesn't exist.

But let me put that argument aside and assume, as Silvia et al. (2008) do, that the constructs of generic creative ability and generic divergent thinking ability reflect actual abilities that people possess in varying degrees. Given this assumption, it is fair to ask whether their proposed scoring represents a possible improvement on current methods of scoring. Unfortunately, Silvia et al. (2008) have failed to present convincing evidence that this might be the case.

John Baer, Teacher Education Department, Rider University.

Correspondence concerning this article should be addressed to John Baer, Memorial 102, Rider University, 2083 Lawrenceville Rd., Lawrenceville, NJ 08648. E-mail: baer@rider.edu

Due to space limitations, I will focus on just two problems with their method and analysis: a misunderstanding (and consequent misuse) of the Consensual Assessment Technique, and a deeply flawed validation process. There are other problems, such as the conceptual problem of confounding divergent thinking and evaluative thinking (because even if the new scoring system did lead to scores that correlated with actual creative performance, it might be caused solely by subjects' evaluative thinking skills, not their divergent thinking skills—that is, it might be a test related to creativity, but not to divergent thinking ability).

Problems Associated With the Proposed Use of the Consensual Assessment Technique

Silvia et al. (2008) write that:

> Subjective scoring of creativity—particularly Amabile's (1982) *consensual assessment technique*—has been popular for several decades in the study of creative products. The consensual assessment technique entails independent judges—ideally but not necessarily experts—rating products for creativity, based on the judges' tacit, personal meanings of creativity. Judges often show high consistency and agreement (Amabile, 1982; Baer, Kaufman, & Gentile, 2004; Kaufman, Gentile, & Baer, 2005; Kaufman, Lee, Baer, & Lee, 2007). (p. 70)

The assertion that judges need not be experts is problematic. The validity of the Consensual Assessment Technique is grounded in and completely dependent on the use of expert judges, and unless novice judges' ratings can be shown to closely match those of judges who are appropriate experts in the domain of the artifacts being judged, their assessments cannot be claimed to have validity. The Consensual Assessment Technique works the way real-world creativity assessment works—one asks experts to use their independent and unfettered domain-based expertise to judge the creativity of a set of artifacts, basing their ratings on comparisons within the set of artifacts being judged. In doing so psychologists must totally put aside their own theories about what is creative and rely on the experts' unhampered assessments (or as Silvia et al. [2008] put it in the paragraph quoted above, "based on the judges' tacit, personal meanings of creativity"). When a Nobel Prize winner is chosen, they don't apply a rubric that psychologists have created. Nor do people outside the field get a vote. It's decided by a consensus of experts in the domain in question. This method may not be perfect—it certainly isn't perfect—but it's quite simply the best possible assessment of creativity at a given time (ideas about creativity in domains may change over time, of course) and in a given domain. This is the way the

Consensual Assessment Technique works, and its validity requires judges who possess domain-based expertise (or some way to validly mimic the judgments of such experts). If psychologists or anyone else should happen to disagree with the consensus of experts in such fields as poetry, cosmology, sculpture, or history about which are the most creative accomplishments in their respective fields, one can only conclude that the psychologists are wrong. It is the experts in a domain who define what in their domain is creative, what is accomplished, and what is neither.

When using the Consensual Assessment Technique, panels of judges are assembled based on the nature of the artifacts to be judged. If the artifacts whose creativity is to be judged are poems, one assembles a group of poets, each working independently. If the artifacts are collages, one gets artists to judge their creativity. Just as in major prizes, experts in the domain make the judgments. They tend to agree, resulting in high interrater reliabilities, and their judgments of creativity are also not the same as their judgments of other things, such as technical goodness, neatness, or expressiveness (Amabile, 1982, 1996; Baer, 1993; Baer et al., 2004; Kaufman et al., 2005).

It is of course possible that novices might give similar judgments as expert judges in some tasks, but that is an empirical question that has not been clearly answered. To the extent that it has been answered, however, it appears that the answer is that novices and judges, while both tending to agree among themselves, provide different ratings (Kaufman, Baer, Cole, & Sexton, in press). Amabile discusses "Who Are 'Appropriate' Judges" (1996), but concludes that "it would be a mistake to conclude that everyone (or even every psychology graduate student) can be considered an appropriate judge" and "the best guideline is to use judges who have at least some formal training and experience in the target domain" (pp. 72–73).

Amabile reported that in some domains, psychology graduate students and teachers gave somewhat similar creativity ratings (with *r*s of 0.44, 0.65, 0.69, and 0.80) as experts when rating artifacts produced by children (and one might argue that part of the expertise one needs to judge, say, the creativity of stories written by young children is a familiarity with children's writing, such as teachers might be expected to have, together with some level of expertise in the area of fiction). But the highest (0.80) such correlation Amabile reported was not between experts and novices; it was the correlation between the ratings of poets, who are indeed experts in the domain of poetry, and English literature graduate students and senior honors students, who simply have a somewhat different kind of expertise in the domain. The 0.44 and 0.65 *r*s were from a study that compared members of Stanford's psychology department, art teachers, and artists; and the 0.69 correlation was between seven artists and seven nonartists described as psychology graduate students, undergraduates, and elementary school teachers (she doesn't say how many of each in the total of seven "nonexperts"). So Amabile was really talking about different *kinds* of experts in most cases (not a comparison or experts and nonexperts), and in other cases comparing quasi-experts (psychology graduate students judging the work of young subjects' collages) and experts. And even then, the correlations were not that impressive. Showing that they have 20% to 45% of shared variance certainly isn't saying that their ratings are the same. Levels of agreement such as these might be sufficient for a large research program using only group comparisons, but they are too low to be acceptable for any use that involves individual score reports.

Similarly, no real novices were involved in the Baer et al. (2004) and Kaufman et al. (2005) studies that Silvia et al. (2008) cite; the nonexperts were all at least quasi-experts in each case. For example, Kaufman et al. (2005) demonstrated that gifted adolescent creative writers (who had been selected for a statewide summer writing program for highly gifted writers) gave similar creativity ratings as adult experts ($r = 0.78$ for poetry creativity ratings and $r = 0.77$ for short story creativity ratings).

Silvia et al. (2008) are not the first creativity researchers to make the mistaken claim that experts are not needed when using the Consensual Assessment Technique. Kasof, Chen, Himsel, and Greenberger (2007) recently suggested that "Undergraduate students have been found to provide reliable and valid judgments of creative products" (p. 115) and quoted Amabile (1983), who wrote that "there is no clear superiority of artists over nonartists in average interjudge correlations" (p. 57). But as noted earlier, Amabile was actually comparing ratings of experts with those of other judges with middling levels of expertise—quasi-experts. She was not comparing experts and novices. And in saying that "there is no clear superiority of artists over nonartists in average interjudge correlations," Amabile was only talking about levels of interjudge correlations—which only assess reliability, not validity. Validity, for the Consensual Assessment Technique, is contingent upon the use of expert judges. Showing that both expert judges tend to agree with other expert judges and that nonexpert judges tend to agree with other nonexpert judges—that is, that each group has good interrater reliability within its own group—can tell us nothing about how well the ratings of experts and nonexperts match. It is rather like comparing Directors' Guild Awards and Peoples' Choice Awards, both of which might achieve good interrater reliability but nonetheless yield different judgments. Would this allow one the claim that People's Choice Awards are equally good assessments of quality? I think not. Only if novices can be shown to give nearly similar creativity ratings as experts—something that has manifestly not been demonstrated by Silvia et al. (2008)—can one assume that it doesn't matter whether one uses novices or experts to rate creativity.

There is a valid question regarding who might be the appropriate experts for judging the creativity of responses to a divergent thinking test. I'm not sure of the answer (or if there is an answer): It may be that psychologists who study divergent thinking could be appropriate judges, and perhaps an argument could even be made that the appropriate judges *are* novices (in the same way that one might argue that People's Choice Awards are validly measuring a kind of quality, even if it is different kind of quality from that measured by Directors' Guild Awards).[1] But Silvia et al. (2008) make no such claim of expertise for their judges, instead arguing that experts are

[1] As an assessment of quality, most would probably agree that Directors' Guild Awards are generally better indicators, but in predicting which films will have better box-office success, the People's Choice Awards might be more accurate. Both can be valid measures, but of somewhat different things.

simply not needed; in fact, they tell us little about the judges, only about the instructions they were given. (There were serious problems with the instructions also, as will be explained below.) The important point is that for the Consensual Assessment Technique to be a valid measure of creativity, one needs to present convincing evidence that the judges do in fact possess the appropriate expertise. Simply saying that novices might do just as well does not negate this requirement, and unless one can show that the judges had the appropriate expertise (or can be shown to reliably mimic the judgments of those who possess that expertise), use of the Consensual Assessment Technique is not valid.

Then there is the question of instructions to raters. Silvia et al. (2008) wrote "agreement between raters can be enhanced by giving them clear instructions, by providing accepted definitions of creativity, and by training them in the scoring system" (p. 71). This is certainly true, but it undermines the integrity of the Consensual Assessment Technique. (Giving clear instructions to judges can also virtually guarantee high interrater reliability, but then such a measure means nothing more than the ability of the judges to follow precise directions.) Amabile insisted that judges work independently and that it is their judgment, not the experimenters' judgments, that must be accepted. She wrote:

> The essence of the consensual definition is that experts in a domain can recognize creativity when they see it. . . If experts say (reliably) that something is highly creative, we must accept it as such. The integrity of the assessment technique depends on agreement being achieved without attempts by the experimenter to assert particular criteria. . . Thus, the judges should not be trained by the experimenter to agree with each other [and] they should not be given specific criteria for judging creativity. (1983, p. 38)

Silvia et al.'s (2008) training procedures directly violate the principles upon which the validity of the Consensual Assessment Technique is based. This is a fundamental part of the consensual assessment technique, not something that can be incorporated or ignored depending on the wishes of the researchers. Silvia et al. (2008) themselves acknowledged this when they wrote that "The consensual assessment technique entails independent judges. . . rating products for creativity, *based on the judges' tacit, personal meanings of creativity*" [italics added]). And yet Silvia et al.'s (2008) procedure directed the judges to ignore (or supersede) their "tacit, personal meanings of creativity" by giving them "clear instructions, by providing accepted definitions of creativity, and by training them in the scoring system."

There is one more problem related to using a variant of Amabile's Consensual Assessment Technique for scoring divergent thinking tests that would make Silvia et al.'s (2008) scoring system problematic, even if they (a) used expert judges (or showed that novices gave similar ratings) and (b) stopped training judges to give the kinds of ratings the experimenter wants rather than rely on their own expert judgment (which is the soul of the Consensual Assessment Technique). The problem is that the Consensual Assessment Technique can provide excellent creativity ratings, but those ratings are dependent upon the particular set of judges and the particular set of artifacts that they judged. One cannot meaningfully compare ratings of different groups of untrained experts on different sets of artifacts. When using the Consensual Assessment Technique judges are making *comparative* ratings, not absolute ones, within a limited sample; a poem that might be judged highly creative in a sample made up of very pedestrian poems might be judged to evidence little creativity in another sample that includes a wealth of highly original poems. There is no way to standardize ratings with the Consensual Assessment Technique. This means that even if Consensual Assessment Technique-based ratings of divergent thinking prompts could be made to work (such as in research studies, which is where the Consensual Assessment Technique has mostly been employed), they could still not produce meaningful standardized scores, scores that could be compared with other divergent thinking test scores produced using the same technique but with different samples and raters. It boils down to this: One cannot have it both ways. A researcher can use the Consensual Assessment Technique to rate creativity within a sample very effectively (reliably and validly), but doing so precludes giving raters the kinds of directions one must give if one is to establish an independent rating scale that isn't dependent on the particular sample of artifacts being judged. (Conversely, one can fairly easily get standardized ratings by giving judges very precise instructions on how to evaluate each artifact, but these ratings cannot claim any validity based on the expertise of the judges. Whatever validity such ratings might have would come from the rubric they were given; it is therefore only the judgment and skill of the psychologist who created the rubric that matters in such cases. But there is a good reason why psychologists are not typically recruited to evaluate works in other domains. They generally lack the requisite expertise.)

Problems With the Validation Procedure (Study 2)

Silvia et al. (2008) used two kinds of measures for their validity assessment: personality measures and choices of college majors, both of which they argue are associated with creativity. I believe the connection between the personality measures they have used and creativity is, at best, tenuous, and adding yet another link to an already questionable chain of connections in order to make a validity claim is the kind of stretch that only a true believer would be willing to make. As such this approach provides little assurance that their measure is actually assessing creativity in a meaningful way. It would have been far more convincing if they had linked their divergent thinking task ratings to creativity ratings of actual creative products (e.g., by having subjects write stories and/or poems, make collages, etc., and using the Consensual Assessment Technique to assess the creativity of those artifacts). Although their college student subjects were indeed "too young to examine the relationship between creative accomplishments across the life span and divergent thinking," it is not necessary to wait to see if their subjects become geniuses in the next half century in order to assess their creativity. One can meaningfully assess the creative performance of college students, as Silvia et al.'s (2008) endorsement of the Consensual Assessment Technique acknowledges, and such a validation attempt would have been much more convincing than one based on uncertain relationships between creativity and certain personality configurations.

Even more problematic is the use of college majors as indicators of creativity, which evidences a very impoverished conception of creativity. Silvia et al. (2008) write that students who are "pursuing arts majors—majors devoted to the fine arts, performing arts, or decorative arts—have chosen to devote their college years to

receiving training in an artistic field, and training is necessary for later creative accomplishment" (p. 77). Students choosing these majors score 1 for creativity; students choosing any other major, which presumably is both a sign that they lack creativity in the present and a choice *not* to prepare for later creative accomplishment, score zero for creativity.

By this thinking, students majoring in art history are generally more creative than those majoring in mathematics, those majoring in art education are more creative than those majoring in science education, students majoring in apparel products design are more creative than those majoring in psychology, and those majoring in graphic design are more creative than those majoring in English. Silvia et al. (2008) seem to be claiming that (a) students studying anything related to the fine or performing arts (or teaching in these areas) are both more creative now and are also forming more of a foundation for future creative work than students majoring in the sciences, social sciences, or humanities, and (b) those students who have presumably wasted their college years learning chemistry, or history, or philosophy, or mathematics are not only lacking in creativity at present, they are also doing little that might constitute a foundation for later creativity.

I am not at all sure what it is that Silvia et al. (2008) are measuring using their scoring method, but their own data suggest that it is *not* generic creativity. Based on the association they report between the divergent thinking scores produced by their method of scoring and college majors in the arts versus college majors in other areas, it appears that what they are measuring is something related to an interest in the arts—possibly related to divergent thinking skill in the general thematic area of the arts or in some domain within that general thematic area, although it is impossible to know this based on the data they report (see Baer & Kaufman, 2005, and Kaufman & Baer, 2005, for a discussion a hierarchical model of creativity that would allow for factors at varying levels of generality that might impact creativity in diverse domains, in single domains, or only in subdomains).

This brings me back to where this essay began—this issue of domain specificity. I acknowledge that the generality-specificity issue is unresolved, and it may be true that general, all-purpose, domain-transcending creativity thinking abilities may exist (which would mean, for example, that a creative musician could, if he chose to do so, apply some of what makes him creative in music to accounting or teaching and to be more creative thereby; general intelligence might be one such domain-transcending kind of ability that could impact creative performance in almost any field). If such generic creative thinking skills, traits, or dispositions exist, even in the very limited sense that a hierarchical model of creativity might propose, it may also be true that people vary in their levels of such creativity-relevant skills, traits, or dispositions. But Silvia et al. (2008) have themselves provided fairly convincing evidence that this is *not* what their test is measuring, because their evidence implies that it is measuring abilities more relevant to some general thematic areas or domains (something having to do with interest in the arts) than others.

I am uncertain whether or not divergent thinking testing is or can be useful. It may be that a new divergent thinking testing or scoring procedure will one day overcome some of the problems of current divergent thinking tests that Silvia et al. (2008) have noted. It is also possible, of course, that continued failure of attempts to improve divergent thinking testing may lead to a general reassessment of the potential value of divergent thinking tests as predictors of creativity. A test of divergent thinking that could reliably and validly predict creative performance would indeed be a valuable tool. Unfortunately, the evidence presented thus far for Silvia et al.'s (2008) proposed method for scoring responses to divergent thinking tasks has far too many flaws to allow any confidence in its use. It is to be hoped that if they or others wish to pursue this method in the future they will try to use more meaningful ways to check the validity of their results.

References

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

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity.* Boulder, CO: Westview.

Baer, J. (1991). Generality of creativity across performance domains. *Creativity Research Journal, 4,* 23–39.

Baer, J. (1993). *Divergent thinking and creativity: A task-specific approach.* Hillsdale, NJ: Erlbaum.

Baer, J. (1994a). Divergent thinking is not a general trait: A multi-domain training experiment. *Creativity Research Journal, 7,* 35–46.

Baer, J. (1994b). Generality of creativity across performance domains: A replication. *Perceptual and Motor Skills, 79,* 1217–1218.

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

Baer, J., & Kaufman, J. C. (2005). Bridging Generality and Specificity: The Amusement Park Theoretical (APT) Model of Creativity. *Roeper Review, 27, 158*–163.

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

Kasof, J., Chen, C., Himsel, A., & and Greenberger, E. (2007). Values and creativity. *Creativity Research Journal, 19,* 105–122.

Kaufman, J. C., & Baer, J. (2005). The amusement park theory of creativity. In J. C. Kaufman & J. Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 321–328). Hillsdale, NJ: Erlbaum.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (in press). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal.*

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

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

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

Received October 30, 2007
Revision received October 30, 2007
Accepted November 2, 2007 ■