JAMES C. KAUFMAN
JOHN BAER
JASON C. COLE

Expertise, Domains, and the Consensual Assessment Technique

ABSTRACT

The Consensual Assessment Technique (CAT) argues that the most valid judgments of the creativity are those of the combined opinions of experts in the field. Yet who exactly qualifies as an expert to evaluate a creative product such as a short story? This study examines both novice and expert judgments of student short fiction. Results indicate a need for caution in using non-expert raters. Although there was only a small (but statistically significant) difference between experts' and novices' mean ratings, the correlation between the two sets of ratings was just .71. Experts were also far more consistent in their ratings compared to novices, whose level of inter-rater reliability was potentially problematic.

INTRODUCTION

The Consensual Assessment Technique (CAT) is one of the most powerful tools that researchers have to assess creativity. The CAT was initially developed and validated by Teresa Amabile (1982) and further developed by her and other researchers in the last quarter century (Amabile, 1983, 1996; Baer, 1993, 1994a, 1994b; Baer, Kaufman, & Gentile, 2004; Baer & McKool, in press; Hennessey, 1994; Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Lee, Baer, & Lee, 2007). Because it uses assessments of real products (not test items manufactured by psychologists based on a theory of creativity) that are judged in the same way creativity is judged in the world outside the laboratory — using panels of actual expert judges in the domain in question — the CAT has been called the "gold standard" of creativity assessment (Carson, 2006). It can be used in virtually any domain, and it is not linked to or dependent for its validity on any particular theory of creativity.

The CAT is based on the idea that the best possible measure of the creativity of a work of art, a theory, or any other artifact is the combined assessment of experts in that field. Whether one is selecting a poem or a work of fiction for a prestigious award or judging the creativity of a sculpture or a scientific theory, one doesn't compute a creativity rating by following a checklist or applying a general creativity-assessment rubric. The most valid possible judgments of the creativity of such works or theories at any point in time — imperfect though these

assessments still may be — are the combined opinions of experts in the field. That's essentially how most prize committees work, and this is why only the opinions of experts matter when choosing, for example, the winner of the Nobel Prize in chemistry. The opinions of the average person on the street do not matter at all. The Consensual Assessment Technique uses essentially the same procedure to judge the creativity of less exalted artifacts, such as collages, stories, or poems produced by research subjects. These are rated for creativity by experts in that domain (e.g., artists or art teachers for collages, and fiction writers, editors, or creative writing teachers for short stories).

Such ratings are necessarily comparative. Experts rate the artifacts they are given in comparison to other artifacts in the same group, not to any absolute standard. Although experts make their judgments privately and independently based solely on their own sense of what is creative in their field with no input or instructions from the researchers and no discussion of any kind with other experts about the task, these assessments tend to produce quite good inter-rater reliability, generally in the .80 to .90 range, although this varies with the number of judges — the more judges, the higher the inter-rater reliability (Amabile, 1982, 1996; Baer, 1993; Baer, Kaufman, & Gentile, 2004).

The CAT is not without limitations, however. Because it relies on comparisons of levels of creativity *within a particular group*, the CAT does not produce any kind of standardized scoring system that might allow comparisons to be made across settings. Its widest use to date has therefore been in research. There is a limited possibility of cross-group assessment and comparison, which can be achieved by including a standard subset of items within several discrete groups of artifacts that are being judged separately (Baer & McKool, in press), mirroring the way that test developers use the equating technique of placing shared items in different forms of a test to show equivalence across those different forms of the same test. But full standardization of the CAT is not possible.

Similarly, judgments of creativity, especially at the highest level, change over time (Csikszentmihalyi, 1999). There is no eternally perfect assessment of creativity, in part because all judgments of creativity are constrained by the zeitgeist and the evolving standards of each domain in which creativity is assessed at the moment such assessments are made. This is true both of assessments at the highest, Big-C levels of creativity (Nobel Prizes, Fields Medals, Pulitzer Prizes, etc.) and at the everyday, little-c levels of creativity that are the usual focus of CAT creativity assessments. However, such changes of standards within domains are not the same as disputes among creativity researchers regarding which theory of creativity is better. The best and most appropriate judges of creativity in a domain at any point in time are the experts in that domain at that point in time, regardless what psychologists might think about creativity. Changes in what psychologists think about creativity matter (in terms of both real-world and CAT-based creativity assessment) *only* if they influence the opinions of experts in some domain (which is probably rare) or in some way mimic what experts in a given domain judge to be creative.

A key element of the CAT, therefore, is the assessment of creativity *by experts in the domain in question*. It would make no sense to have engineers judge the creativity of poems, poets judge the creativity of cake decorations, or chefs assess the creativity of engineering designs; for the CAT to be valid, judges need

to be experts, not novices in a domain. It is this mooring in the judgment of experts that underwrites the validity of the CAT. But what constitutes an expert? Amabile (1996) grants that some tasks in some domains may require only limited expertise and training in a domain. However, she advises 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" (p. 72). In fact, "some formal training in the field may be necessary for judges even to understand the products they are assessing" (p. 72)[1].

Experts are neither as numerous nor as readily available to serve as judges as nonexperts, however, and there has therefore been a temptation among creativity researchers to substitute judges who are somewhat short on credentials as actual experts in a domain. For example, Baer (1996a), who is one of the authors of this study, once used preservice elementary school teachers as judges of the creativity of collages created by young children. Other researchers have also used nonexpert judges (e.g., Chen, et al., 2006; Joussemet & Koestner, 1999; Kasof, Chen, Himsel, & Greenberger, 2007; Niu & Sternberg, 2001). It is not clear exactly what kind of expertise one needs to be an expert rater in a given domain. While acknowledging that preservice teachers are not experts in the domain of collage design (as in the Baer, 1996) study just cited, just who *are* the most appropriate experts? Following Amabile's (1996) advice, might preservice teachers' experience with young students' artwork be sufficient? It would certainly be better to have artists (especially artists with some familiarity with children's artwork) or experienced art teachers as judges. But might the quasi-expertise of preservice teachers be good enough?

Probably the only way to answer this question would be actually to compare the ratings of experts in the domain with ratings of what we have termed quasi-experts to see how well they match up. And what about novices — judges who have no explicit training or expertise at all in the domain of the artifacts to be judged, judges who cannot even be termed quasi-experts. While certainly not qualified to judge candidates for the Fields Medal or the Booker Prize, might novices be good enough as judges at the little-c level of judging student artwork, poems, or stories? These are empirical question, and as Amabile (1996) suggests, the answer might vary from domain to domain. If novices' judgments in a given domain matched those of experts, then such a finding would be a boon for creativity research (because assembling panels of novices is easy, whereas experts are in short supply). But if novices' judgments do *not* match those of experts, then the use of novice judges may produce untrustworthy and unreliable results. In such a scenario, researchers should *not* use novice judges, despite their ready availability.

There are a handful of studies that have compared novice and expert judgments. Runco, McCarthy, and Svenson (1994) reported that expert assessments of artwork may be harsher than peer or self assessments. Unfortunately, the number of expert judges ranged from only one to three in this study, which did not allow assessments of inter-rater reliability or sufficient data on how expert and

[1] It is important to distinguish expert ratings used for the CAT from self-ratings used to analyze people's self-perceptions of their own creativity (e.g., Kaufman, 2006). In such cases, these "non expert" ratings are not designed to measure genuine originality or quality.

novice ratings compared. Kaufman, Gentile, and Baer (2005) compared the ratings of expert judges and gifted novices (in this case, gifted high school students who were highly interested and talented in the domain being rated). The creativity ratings made by these gifted novices evidenced good inter-rater reliability and were significantly correlated with the creativity ratings of actual experts in those domains. This at least partially opens to door to the use of nonexperts, but gifted novices are not at all the same as nonexperts, and whether they are more like experts or more like novices is unclear. They probably fall somewhere in between the two groups, suggesting that there is perhaps a continuum stretching from true experts at one end to complete novices at the other, with various levels of partial or quasi-expertise in between.

Amabile (1996) compared the creativity ratings made by experts and by people we are calling quasi-experts, people similar to the gifted novices of the Kaufman et al. (2005) study. In one of Amabile's studies, three groups of judges rated the creativity of a small collection of collages: a group of psychologists, a group of art teachers, and a group of artists. The latter two groups have the kind of expertise Amabile argued was essential ("judges who have at least some formal training and experience in the target domain"; Amabile, 1996, p. 73), while the group of psychologists lacked such training but might, because of their knowledge of children (and perhaps of creativity research), be thought of having at least some related expertise. The correlation between the ratings of the artist-judges and the psychologist-judges was just .44. While the psychologists lacked artistic expertise, they did have a different type of expert knowledge (i.e., understanding children) that might have been relevant to making these judgments and thus they cannot be considered complete novices or nonexperts. There were nonetheless rather large differences in how the judges with artistic expertise and the psychologists rated the creativity of the collages. The .44 correlation between the artist-judges and the psychologist-judges hardly suggests that experts' judgments could simply be replaced by that of the psychologists. Although this was a relatively small study, it cautions against the use of nonexpert raters — even ones with some related experience that might qualify them as quasi-experts — in rating the creativity of children's collages.

Dollinger and Shafran (2005) used a significantly modified version of the CAT and found the expert judges and novices produced fairly similar ratings. Twenty participants' drawings were judged by five artists (experts) for "quality of drawing" and "overall creative Gestalt" (p. 595). Five novices (or perhaps quasi-experts; all five had psychology backgrounds and some had graduate degrees, although it is not clear if they were students of creativity) also judged the 20 drawings, but they first underwent a brief training activity that resembles the kinds of calibration training given to holistic scorers in other (non-creativity) assessments. However, a few caveats must be noted. First, Dollinger and Shafran (2005) did *not* use the CAT; as they themselves appropriately noted, "any training to calibrate judges violates one assumption of the Consensual Assessment Technique" (p. 593), so this was not a CAT procedure. But if one *could* actually find a way to duplicate the creativity ratings that expert judges would produce without the need to cajole actual experts into participating (novices being both plentiful and cheap), that would nonetheless be a good thing and would yield valid results (after all, they would be the same results that would have been obtained had actual experts

made the judgments). This appeared to be Dollinger and Shafran's goal. They found that trained nonexperts did produce similar ratings to experts (.87 for "quality of drawing" and .90 for "overall creative Gestalt"). This was a small study (just 20 drawings), but it does suggest that training of this kind — which, it should be noted, is not training based on the experimenters' conceptions of creativity in any way, but is instead based *completely* on the opinions of experts about the creativity of similar artifacts — might be able to reduce the need for expert judges. (Unfortunately, their nonexperts were actually quasi-experts, so one cannot say how well true novices could be trained. They also did not include ratings by untrained novices for comparison.) But even if this training can be shown to work, experts are still needed: for training purposes, one must still have a set of artifacts of the same kind that have been judged previously by a panel of experts. For every specific task for which CAT ratings might be desired, an expert-judged set of artifacts would be needed to determine whether the novices had been successfully trained.

It may be possible in a brief training session, and especially in some rather very tightly constrained domains, to get novices to align their judgments on a narrowly defined task to those of a group of experts. But it seems unlikely that a brief training session could turn novices into experts in a domain more generally (if so, one could save apprentices a great deal of time in the training they undergo in most domains to become experts!). And we simply don't know whether such training might work for other tasks commonly used in creativity assessment, such as making collages or writing poems or stories or picture captions, or at what level of talent or creativity evidenced in the artifacts such training would no longer work. It is of course quite unlikely that such training could work at the Big-C level of creativity — Nobel Prize selection committees need not fear replacement by novices — but it remains an open empirical question whether the CAT could be replaced or modified for little-c tasks.

Do novices make ratings that are virtually indistinguishable from the ratings of experts rating the same set of artifacts? If they do, then there is no reason not to allow nonexperts to stand in for experts in making creativity ratings for that particular kind of artifact. But if nonexperts do not give ratings that are highly similar to those of experts, then they simply are *not* valid CAT measures of creativity. These novice-based assessments may of course be valid measures of something other than creativity. By analogy, Peoples' Choice Awards may not be as valid assessments of creativity as, say, Directors' Guild Awards, but they may nonetheless be good measures of something else, such as commercial viability. And we emphasize that the use of nonexpert judges, even if shown to be valid for one task, may not work in a different domain or even on a different task in the same domain, so such use must be validated for use with each different kind of artifact.

Kaufman, Baer, Cole, and Sexton (2008) recently put this question to a direct empirical test in the domain of poetry by having 10 expert and 106 nonexpert raters judge the creativity of 205 poems. The poems had been written by college students, and the nonexpert raters were other college students (not the same students who wrote any of the poems in the study). The novices' creativity ratings were found to be quite different than those of the experts, which suggests that substituting novices for experts, at least in the domain of poetry written by

college students, simply won't work. Nonexpert raters' judgments of creativity were also found to be inconsistent (showing low inter-rater reliability), unlike those of the experts (which, as in past studies, were found to yield quite good inter-rater reliability).

Evidence that novices do not rate the creativity of poems in ways similar to experts is sound reason to avoid the use of novice judges in creativity studies using poetry, at least poetry written by college students. (Whether novice judges could validly assess the creativity of poems written by children is unknown, but the results with college students' poems certainly suggest caution.) But what of other domains? Creativity has been shown to vary widely across domains, and even across some tasks in the same general domains, such as poetry-writing and story-writing (Baer, 1993, 1996b; Kaufman & Baer, 2005). There is also variability across domains in the kinds of skill and knowledge one needs to be an expert in a given domain; artists, poets, and biologists have very different kinds of expertise. Might there also be variability (as Amabile, 1996, suggested) in the amount of expertise one needs to make judgments similar to those of true experts in a domain? Unfortunately, this question can only be answered by separate studies in each domain in question: there is no general heuristic to guide researchers through the important but difficult decision regarding selection of judges. One can always err on the side of caution and use only judges who clearly have the appropriate expertise, but that is difficult and often expensive. Using novices whose judgments sufficiently match those of experts could make research using the CAT considerably easier, which would be a boon for creativity research; but using novices whose judgments do not match those of experts can only result in misleading and invalid results and conclusions, something no researcher wants to risk. There is therefore a need for studies that compare novice and expert creativity ratings in different domains, especially ones (such as poetry, short stories, and collage-making) that are so widely used in creativity research.

It has been argued that "critical abilities are more advanced than productive abilities" (Perkins, 1981, p. 128), an assessment with which Johnson-Laird (1988) concurred, calling this the "central paradox of creativity" (p. 208). It has also been suggested that domain specificity may affect evaluative thinking in domain-specific ways parallel to the ways domains influence divergent thinking (Baer, 2003). Might the critical abilities of nonexperts be sufficiently "advanced" (to use Perkins's term) to bring their evaluations in line with those of experts, at least in some domains, even though they may not have expertise or be themselves productively creative in the domain? We know that novices are not valid judges of the creativity of poems written by college students (Kaufman, Baer, Cole, & Sexton, 2008) and have reason to doubt that psychologists might replace artists as judges of children's collages (Amabile, 1996). But those are just two domains. What of other areas, perhaps domains in which novices might be expected to have greater familiarity? Most college students have little experience with poetry, and most psychologists have little experience with collages. Might novices' assessments of the creativity of short stories (fiction being a domain with which most novices have some familiarity) more closely match those of experts in the field? It is precisely that question that this study was designed to answer.

For a group of novice judges' rating to be acceptable replacements for the judgments of experts, two things need to be demonstrated:

1. Good inter-rater reliability among the novice judges (because if they don't tend to agree, then whatever their mean ratings may be is essentially dependent on which novice judges happened to be included in the pool).
2. A high correlation between the ratings of the novices and experts in the domain, high enough to supplant experts per standard criteria minimum recommendations of about .90 (Anastasi & Urbina, 1998) if used for individual score reporting or comparisons, with somewhat lower levels acceptable for group comparisons.

The study reported below compared the creativity ratings of short stories written by college students that were made by two groups of raters: a group of experts in the domain, and a group of randomly selected college students not otherwise involved in the study (that is, students who did not write the stories being judged). We looked at both the inter-rater reliabilities of the two groups and how well the nonexperts' ratings matched those of the experts.

METHODS

Participants

The participants who provided the writing samples consisted of 205 college students from two universities, one a private university from the Northeast and the other a public university in the Southwest. Participants took part in the study online for extra credit. The sample included 54 men and 151 women, with a mean age of 24.20 years (*SD* = 8.73 years).

Procedure

The study was conducted online, where participants first read and signed a consent form, and then were given instructions for the task. In writing the story, participants were instructed to select one from the following two titles, (1) 2305 and (2) execution, and then write a short story in 10 minutes. After completion of the study, all student writings were retrieved from the Website and were identified only by the participant's numbers. All writings were printed in separate sheets with participant numbers on the top of each sheet. The participant stories were then prepared to be rated.

Raters

There were two groups of raters: experts and novices.

Expert raters consisted of ten writers. Of the ten, five had MFAs in creative writing and three others had Ph.D.'s in English. All had published some of their own work and all had experience reading the work of student writers. Six were currently college professors (some tenured or tenure-track, others adjunct) at the time they did this rating. The experts never met or discussed their ratings in any way. There were seven women and three men among the group of expert raters.

Novice raters consisted of 106 college students from California State University, San Bernardino, who participated in the study for course credit. Raters who participated in the first part of the study (writing the poems) were not included. The sample included 25 men and 81 women, with a mean age of 21.17 years (*SD* = 6.21 years). Like the expert raters, the novice raters worked independently. All raters received identical instructions.

RESULTS

The novice judges rated a total of 203 stories (two of the 205 stories were blank and were excluded from this analysis). Their mean rating score was M = 4.79, SD = 0.73. The expert raters rated 203 stories and had a mean rating score of M = 4.50, SD = 1.85. An independent means t test was calculated to compare mean ratings for the two groups, t (401) = 2.01, p = .046. Given the respective means, the t test indicated that expert raters rated the stories as slightly less creative relative to novice raters. It should be noted that this difference, while significant, represents a small effect size (d = 0.21). A Pearson correlation was assessed between the expert and the novice raters after appropriate assumption checks were sufficient. Results revealed a significant relationship between the ratings that had an effect size above large: r (200) = .709, p < .001, according to criteria from Cohen (Cohen, 1988).

Coefficient alpha for the expert raters was .924 (95% CI = .909 – .939), and for the nonexpert raters was .925 (95% CI = .911 – .938), which places these alphas in the excellent alpha range. However, when the Spearman-Brown formula is applied to these alphas, the impact of the number of raters per group becomes quite clear. Standardizing the alpha to be set to 10 raters in each group (per the Spearman-Brown formula), alpha for the experts did not change (as they were just ten raters) whereas alpha for the students dropped massively to .533 (95% CI = .518 – .548).

DISCUSSION

Although the average rating from students and experts demonstrated mean values that were significantly different, this difference had a small effect on the mean ratings. Moreover, story ratings showed highly similar variability between students and experts, with a correlation between their scores of .71, or 50% shared variance. Whereas 50% shared variance does not suggest that students ratings can replace expert ratings, it does suggest that student rating of story creativity has some criterion validity — but not enough to supplant experts per standard criteria minimum recommendations of about .90 (Anastasi & Urbina, 1998). As a comparison, the correlation between student and expert ratings showed very little overlap when rating poems (Kaufman, Baer, Cole, & Sexton, 2008): Pearson r = .22, or 5% shared variance. The difference between the correlations (between experts and students on stories vs. between experts and students on poems) is marked with z score difference of 6.64 (p < .001). The weakness of students shown with the current results is shown in their consistency among one another. Whereas experts achieved an excellent internal consistency of .93 (95% CI = .91 – .94), a rater-equalized (with adjustment to the novice raters' internal consistency to equal the number of raters in the expert group) internal consistency for students showed poor consistency of .53 (95% CI = .52 – .55). Summarily, results showed comparatively little difference between experts and students ratings as far as mean rating and shared variance of the ratings. However, experts were far more consistent in their rating among one another compared to students.

This study offers both some practical and theoretical considerations about the nature of creativity. There can never be one simple rule that will decide who are the most appropriate judges in a given domain and for a given population. Award-

winning poets may be the best judges of the Big-C creativity of poetry, but less famous poets or poetry teachers or poetry critics with more experience reading adolescent poetry might be better judges of teenage creativity in this domain.

Given these findings and those of Kaufman et al. (2008), we propose that one way to determine potential novice-expert agreement might be to look at how popular the domain is with the public. Poetry has largely become less popular among the average reader, whereas novels continue to dominate best seller lists. Perhaps one reason why novices and experts disagree on the creativity of works of poetry and are more aligned on their judgments of fiction is that more novices have experience reading fiction. This concept has also been proposed, in a somewhat different context, by Simonton (2004), who argued that an average person would be hard pressed to competently evaluate the creativity of a scientific theorem or mathematical proof.

Future studies could look at multiple groups of experts and multiple domains. How would experts, quasi-experts, gifted novices, and novices all compare? Would patterns of differences be similar in the areas of visual arts, mathematics, or other key domains? These studies are time consuming and potentially costly; however, we believe that such work will be an important step forward in creating a comprehensive creativity assessment across multiple domains.

In the meantime, what guidance do these results afford researchers using the CAT? Finding out how well novices can replace expert judges when using the CAT in a given domain — or what level of quasi-expertise might be needed — is primarily a practical question, but unless one has reason to argue that the creativity ratings of a particular group of novices or quasi-experts will closely match those of experts, it is probably safest to adhere to Amabile's (1983, 1996) advice and use experts whenever possible. Unfortunately there is no simple rule to determine when quasi-experts, or even novices, might validly replace experts in judging creativity in a given domain or situation. Using larger numbers of novices than one would normally need for experts is likely to improve their inter-rater reliability (which rose from an unacceptable .533 with 10 judges to .925 with 106 novices). One must still ensure that these ratings do indeed correlate to a sufficiently high degree with the ratings of experts, however, because the validity of the CAT is grounded in the expertise of the judges. Studies that will include a wider range of creativity (from very low to very high) are also likely to result in both higher inter-rater reliabilities and better matches between novice and expert ratings of creativity.

We agree with Carson (2006) that the CAT is the "gold standard" in creativity research, but researchers using the CAT need to justify the judges they use. If the judges are experts in a domain, that is enough justification. To the extent that judges are either novices or quasi-experts, researchers must make a case for why the judges they have chosen are appropriate. For example, such an argument could be based a previous study that showed a correspondence between expert judges' ratings and those of the kind of quasi-expert or novice being used in the proposed study (provided, of course, that they were judging artifacts of the same kind and produced by subjects from the same population in both the previous and the proposed studies). For some tasks, only expertise of the highest level could be expected to work, while for others (e.g., caption-writing or collage-making by young children), judges with lower levels of expertise or even novices might

be appropriate. The kind of training that Dollinger and Safran (2005) have used, in which novices were trained by examining artifacts and the ratings given by expert judges, might be shown to work in some domains, but of course evidence must be provided that it does indeed work in the particular domain in question. If researchers make their case why they believe the judges they have used are appropriate, readers can decide how much confidence to put in the findings of that research.

Practically, this line of research can enhance a popular classroom teaching tool, collaborative learning. In this technique, students are often encouraged to learn not only from the teacher, but also from each other (Burke, 1994; Elbow & Belanoff, 1999; Mueller & Fleming, 2001). With the advent of computers and the Internet, students can interact with a much larger peer group than ever before (McFadzean & McKenzie, 2001). One viable way of assessing student creativity might be collaborative learning, but this line of research should raise some yellow flags. In order for the evaluations to have any real meaning, the students should have some expertise in the domain they are ratings. In addition, the specific domain chosen for students to demonstrate and critique creative work is important. This study indicates that short stories may represent a better option than poetry — but neither is ideal. Previous work by Kaufman, Gentile, and Baer (2005) demonstrating that gifted students can be quite good judges of creative writing suggests that use of such peer judgments may be especially appropriate when working with gifted students. Future investigations may offer a pattern for solutions to this dilemma of judge selection for both the classroom and the laboratory.

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. (1983). *The social psychology of creativity*. New York: Springer-Verlag.

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

ANASTASI, A., & URBINA, S. (1998). *Psychological testing (7th ed.).* Upper Saddle River, NJ: Prentice Hall.

BAER, J. (1993). *Creativity and divergent thinking: A task-specific approach* . Hillsdale, NJ: Lawrence Erlbaum Associates.

BAER, J. (1996a). Does artistic creativity decline during elementary school years? *Psychological Reports*, *78*, 927-930.

BAER, J. (1996b). The effects of task-specific divergent-thinking training. *Journal of Creative Behavior*, *30*, 183-187

BAER, J. (2003). Evaluative thinking, creativity, and task specificity: Separating wheat from chaff is not the same as finding needles in haystacks. In M. A. Runco (Ed.), *Critical creative processes* (pp. 129-151). Cresskill, NJ: Hampton Press.

BAER, J., KAUFMAN, J. C., & GENTILE, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*.

BAER, J., & MCKOOL, S. (in press). Assessing creativity using the consensual assessment. In C. Schreiner, *Handbook of assessment technologies, methods, and applications in higher education*. Hershey, Pennsylvania: IGI Global.

BURKE, K. (1994). *The mindful school: How to assess authentic learning*. Arlington Heights, IL: IRI/Skylight Training and Publishing.

CARSON, S. (2006). *Creativity and Mental Illness*. Invitational Panel Discussion Hosted by Yale's Mind Matters Consortium, New Haven, CT., April 19, 2006.

CHEN, C., KASOF, J., HIMSEL, A., DMITRIEVA, J., DONG, Q., & XUE, G. (2005). Effects of explicit instruction to "Be Creative" across domains and cultures. *Journal of Creative Behavior*. 39, 89-110.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

CONTI, R., COON, H., & AMABILE, T. M. (1996). Evidence to support the componential model of creativity: Secondary analyses of three studies. *Creativity Research Journal*, *9*, 385-389.

CSIKSZENTMIHALYI, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313-335). Cambridge: Cambridge University Press.

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

ELBOW, P., & BELANOFF, P. (1999). *A community of writers: A workshop course in writing (3rd ed.)*. New York: McGraw-Hill.

JOUSSEMET, M., & KOESTNER, R. (1999). The effects of expected rewards on children's creativity. *Creative Research Journal*, *12,* 231-239.

KASOF, J., CHEN, C., HIMSEL, A., & GREENBERGER, A. (2007). Values and creativity. *Creativity Research Journal, 19,* 105-122.

KAUFMAN, J. C. (2006). Self-reported differences in creativity by gender and ethnicity.*Journal of Applied Cognitive Psychology, 20*, 1065-1082.

KAUFMAN, J. C., & BAER, J. (Eds.). (2005). *Creativity across domains: Faces of the muse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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., BAER, J., COLE, J. C., & SEXTON, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal, 20*, 171-178.

KAUFMAN, J. C., LEE, J., BAER, J., & LEE, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of validity. *Thinking Skills and Creativity*, *2*, 96-106.

MCFADZEAN, E., & MCKENZIE, J. (2001). Facilitating virtual learning groups: A practical approach. *Journal of Management Development*, *20*, 470-494.

MUELLER, A., & FLEMING, T. (2001). Cooperative learning: Listening to how children work at school. *Journal of Educational Research*, *94*, 259-265.

RUNCO, M. A., MCCARTHY, K. A., & SVENSON, E. (1994). Judgments of the creativity of artwork from students and professional artists. *The Journal of Psychology*, *128*, 23-31.

SIMONTON, D. K. (2004). *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge: Cambridge University Press.

James C. Kaufman, Learning Research Institute, California State University San Bernardino

John Baer, Rider University

Jason C. Cole, Consulting Measurement Group