Creativity Research Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hcrj20

Beyond New and Appropriate: Who Decides What Is Creative?

James C. Kaufman [a] & John Baer [b]

[a] Learning Research Institute, California State University at San Bernardino

[b] Rider University

Available online: 10 Feb 2012

To cite this article: James C. Kaufman & John Baer (2012): Beyond New and Appropriate: Who Decides What Is Creative?, Creativity Research Journal, 24:1, 83-91

To link to this article: http://dx.doi.org/10.1080/10400419.2012.649237

CREATIVITY RESEARCH JOURNAL, *24*(1), 83–91, 2012

ISSN: 1040-0419 print/1532-6934 online
DOI: 10.1080/10400419.2012.649237

Routledge
Taylor & Francis Group

Beyond New and Appropriate: Who Decides What Is Creative?

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

John Baer
Rider University

The Consensual Assessment Technique (CAT) is a common creativity assessment. According to this technique, the best judges of creativity are qualified experts. Yet what does it mean to be an expert in a domain? What level of expertise is needed to rate creativity? This article reviews the literature on novice, expert, and quasi-expert creativity ratings. Although current research indicates that novices may be poor choices to be CAT raters, quasi-experts may represent a compromise between ideal scientific rigor and practical time and budget restrictions. Certain guidelines are suggested to make the selection of experts more streamlined, including paying attention to which domain is being assessed.

Believe one who has proved it. Believe an expert. (Virgil)

Many definitions of creativity center on two core elements: novelty and appropriateness to the task or problem being addressed (Amabile, 1983; Baer, 1993; Sternberg, 1999). Mayer (1999) summarized his review of how experts define creativity by saying that "there appears to be consensus that the two defining characteristics of creativity are originality and usefulness" (p. 450). There is an often unacknowledged question that is implicit in such definitions, however: Novel and appropriate (or original and useful) to whom? Who is an appropriate judge of a creative product's novelty and appropriateness?

Novelty may seem fairly straightforward—Is it new? Is it original?—but even among truly creative products and ideas, there are many shades and degrees of novelty, such as those described by the Propulsion Model (Sternberg, Kaufman, & Pretz, 2002), in which creative contributions may range from minor variants of earlier work through paradigm-shifting novelty that changes entire fields of study or endeavor dramatically. Among more everyday creativity, the question of who is an appropriate judge of novelty looms even larger. What might seem very original to someone outside a domain might seem totally pedestrian to someone with expertise in that area.

The question of who defines how appropriate to the task a new idea or product might be is even murkier. Unless there is a clearly established standard or set of criteria, determining how well a product or idea meets the (generally ill-defined) constraints of the task or problem is far from simple. Such criteria are rare; the very nature of creativity is such that it is expected to include the unexpected.

If one looks beyond the world of creativity research, the answer to this "According to whom?" question is fairly consistent. How is creativity most often judged in the real world? By relevant experts. At the highest levels, these might be Nobel Prize, Fields Medal, or Pulitzer Prize committees. More local variants of expert committees are also common (e.g., the scientists and science teachers who are often called upon to judge science fairs or the artists, art critics, and art gallery owners who decide what will be shown at galleries and who write commentary about those pieces).

When it comes to judging real-world creative products, few people look to divergent-thinking test scores,

We thank Dean Keith Simonton for his helpful suggestions.
Correspondence should be sent to James C. Kaufman, Department of Psychology, Learning Research Institute, California State University at San Bernardino, 5500 University Parkway, San Bernardino, CA 92407. E-mail: jkaufman@csusb.edu

psychologist-defined scoring rubrics, or self-assessment checklists. They ask experts. Not everyone will agree with every expert opinion (most years the announcement of the Nobel Prizes in Literature is greeted with outrage or puzzlement). Yet there is no higher court of appeal (except, of course, to other experts within the domain, perhaps with better credentials). There is, perhaps, some irony in this paradox: How could we expect experts to judge those creations that might be changing the very rules that helped establish their own standing in their field?

At the very highest levels of paradigm-shifting creative genius, there certainly are limits to what can be expected of experts—it may sometimes take a new generation to recognize the very greatest creative achievements that are major advances in a domain. There is even a term for this phenomenon: Planck's Principle. Hull, Tessner, and Diamond (1978) studied the ages of both early accepters and continued rejecters of Darwin's *Origin of Species.* Based on these results, they argued that younger scientists were more likely to accept new ideas than older scientists, although a subsequent study (Levin, Stephan, & Walker, 1995) found contradictory results.

Even in such very special cases, domain experts are usually the best choice, even if they may not agree during such revolutionary periods (Kuhn, 1970). At the level of more everyday, garden-variety creativity (the kind being assessed in most creativity research), however, such paradigm shifts rarely come into play. There are also elements to consider beyond mere expertise. Hood (1973) suggested that judges who were very creative themselves were more restrictive in their ratings. Caroff and Besançon (2008) found opposite results; judges who were more original gave higher ratings to more original products.

In these kinds of judgments, it is, of course, important that the experts agree in their evaluations. If different experts came to different conclusions regarding the creativity of a group of products they were evaluating, one could not know whose judgments to trust. But getting agreement among judges—getting good interrater reliability—is only part of what is needed for these evaluations to be valid. Reliability is essential, but it does not in any way guarantee validity. That is why most esteemed prize committees rely on experts. Nonexperts might come to a consensus that the best movie of the millennium was *Twilight*, but what would that mean? Getting judges to agree can only ensure reliability. But if expert judges—the gatekeepers of a domain—agree on what is creativity, then there is evidence of both reliability and validity. It is, of course, possible that experts in a different time may come to different conclusions, but this only shows that fields are not static. Their standards and their paradigms change over time (see Csikszentmihalyi, 1999). But at any point in time, the most valid judgments of the creativity of any product or idea in a domain are the collective opinions of those people who the world has deemed experts in that domain. They define, through their collective wisdom, what is creative and what is not for their domain.

THE CONSENSUAL ASSESSMENT TECHNIQUE: HOW EXPERTISE UNDERWRITES VALIDITY

The basic question of whether expert judges agree on creativity has been explored for nearly a century under the name *aesthetic judgment* (Cattell, Glascock, & Washburn, 1918; Child, 1962). Amabile's (1982, 1983, 1996) Consensual Assessment Technique (CAT) was the first systematic way to demonstrate under what conditions expert judges can be best be used. She answered the question "Creative according to whom?" very directly: Creativity is judged by panels of experts in the relevant domain. Such experts should work independently of one another and are given no guidance for how to rate, other than their own acquired sense of what is more or less creative. She thus followed the way in which real-world creativity is judged, which is perhaps why the CAT has been called the "gold standard" of creativity assessment (Carson, 2006). Panels of experts judges, working independently and without direction, tend to agree which of a group of poems, collages, stories, etc., are the more and less creative. The validity of the CAT is grounded in its use of experts as judges (Kaufman, Plucker, & Baer, 2008).

Use of expert judges certainly limits possible answers to the "Creative according to whom?" question (and the implied "Novel and appropriate to whom?" question that underlies the "novel and appropriate" definition of creativity), but it does not provide a complete answer. What does it mean to be an expert in a domain? Expertise research suggests that it takes approximately 10 years from someone first entering a field to that person making any kind of substantial contribution (Bloom, 1985; Ericsson, Roring, Nandagopal, 2007; Hayes, 1989). These 10 years are spent learning the mechanics of the field, discovering all of the practical issues that can't be taught in a book, and obsessively practicing. These 10 years do not represent a basic apprenticeship, in which one might be taught how to do a trade, such as becoming a tailor. Rather, these are years of active experimentation and new ideas (Gardner, 1993). Reaching the level of greatness may require another 10 years (Kaufman & Kaufman, 2007; Simonton, 2000).

In judging creativity at the highest levels, jurors are expected to have a record of accomplishment in the field in question. Jurors for the Pulitzer Prize in Poetry, for

example, regularly include past winners, renowned English professors, and poet laureates. Yet such an esteemed panel would not necessarily be ideal for judging the creativity of the poems of fourth-graders. In this scenario, a lesser-known expert with experience with children would likely be a preferred choice. In addition, the domain called "poetry" may have many microdomains. Haikus and sonnets are both poems, yet an expert in Haiku may not be the best choice for judging sonnets. Notice also that poetry experts, even at the highest levels, cannot be assumed to have expertise in other domains. The Pulitzer Prize committee for drama would have a very different set of qualifications.

Expertise in the larger world is generally defined by rather narrow domain-based qualifications—one can hardly be a domain-general, all-purpose expert—yet even within clearly defined domains, there are microdomains that may require different types of expertise (see Baer & Kaufman, 2005). An expert in Chaucer may also be an appropriate expert to judge modern poetry, but not if her expertise is mostly limited to Middle English. Similarly, there are also creative products where such expertise is much harder to define. Who would be the appropriate experts to judge the creativity of, say, automobile commercials? People who work in advertising developing such commercials? Car manufacturers? Network executives? Prospective buyers? Filmmakers?

The qualifications of appropriate judges might depend on the goal of the contest in question. The National Board of Review, the Academy Awards, and the People's Choice Awards use very different kinds of judges with different kinds of expertise to assess the same basic product (recently released movies). If one hopes to predict which movies will influence future movie-making, then one might rely on the Academy Award's selections. If one wants to predict which movies are the most critically acclaimed, the National Board of Review typically reflects these views. But if one wants to predict which movies will make the most money, the People's Choice Awards might be a better guide.

WHY NOT USE NONEXPERTS AS JUDGES?

Returning to the assessment of creativity, the use of experts as an answer to the "Creative according to whom?" question often allows one to define potential judges, making Amabile's CAT possible. This definition is not the easiest one, because experts in a domain are often limited, hard to find, or expensive. It can even be cumbersome to obtain judges with a lesser degree of domain expertise, such as graduate students, classroom teachers, or creativity researchers. Such judges are certainly more plentiful than award-winning writers, and research has shown them to have good interrater reliability (Amabile, 1996; Baer, 1993, 1997, 1998; Baer, Kaufman, & Gentile, 2004). But engaging 10 such judges to read and rate 50, 100, or 200 short stories still requires far more resources than giving a group of students a divergent-thinking test. As a result, some researchers have been tempted to use nonexperts as judges of creativity—novices, such as college students with no other qualifications than a desire to earn research participation credit in Introductory Psychology (e.g., Baer, 1996; Chen, Himsel, Kasof, Greenberger, & Dmitreiva, 2006; Joussemet & Koestner, 1999; Kasof, Chen, Himsel, & Greenberger, 2007; Niu & Sternberg, 2001; Silvia, 2008). These novices sometimes agree with one another sufficiently well that the researcher can report adequate interrater reliability (which is easier to achieve with a larger number of judges), but as noted, high interrater reliability alone does not assure validity.

The use of nonexperts as judges can, therefore, be problematic. The validity of the CAT is grounded entirely in the use of appropriate experts as judges. If nonexperts and experts do not agree with each other, then the opinions of experts in a domain should trump those of anyone else.

Many creativity researchers (ourselves included) would often prefer that expertise, a scarce resource, were not required for the CAT to be valid. It would make our work much easier. There is, therefore, a natural temptation to use nonexperts instead of actual experts. But consider where this decision leads. One might poll the students in a middle school regarding movie quality, or invite monolingual judges to rate the quality of films in another language (without subtitles). It is possible (although uncertain; to our knowledge, no research of this kind has been done) that such judges might tend to agree with one another. Yet even if they did, would anyone argue that these are valid judgments of film quality? They might be indexes of appeal to certain audiences (and might even be useful in deciding which movies to promote on, say, the Disney Channel, or which films might be more likely to sell well in other countries). But would they be valid measures of movie quality? Hardly.

WHAT DO WE KNOW ABOUT THE JUDGMENTS OF EXPERTS AND NOVICES?

The use of novice CAT judges might be valid if it could be shown that judges at all spectrums of expertise tended to agree in their creativity ratings. As much as we may wish that novices and experts agreed on assessing creative work, the research does not endorse this position. Regardless of the presence or lack of reliability, studies generally show that novices and experts do not agree. Hickey (2001) conducted an extensive study of novice

and expert ratings of children's musical compositions. Her three composers did not agree with each other (and their ratings could therefore not be used), but she did get agreement for theorists, three types of teachers (instrumental, mixed, and general/choral), and samples of 2nd and 7th grade children. The three types of teachers agreed with each other and with the music theorists. The two groups of children agreed with each other. However, the children's ratings did not correlate with either the theorists' or the teachers' ratings.

Lee, Lee, and Youn (2005) applied generalizability theory techniques to expert and novice ratings of flower designs. Their experts were professional artists who worked in flower design and their novices were undergraduate students. They found low levels of interrater reliability in the novices, as in past studies. They also calculated that the variance due to raters was much less for the experts, also indicating a higher level of agreement. Finally, Lee et al. (2005) found that product-based variance was twice as high in experts as in novices. In other words, novices were much less likely to be able to discriminate between different types of flower designs.

We investigated this relationship in two studies with large samples. We started by asking more than 200 college students write to poems and short stories (for Kaufman, Niu, Sexton, & Cole, 2010). We then gave each set to two different groups of judges: experts and novices. There were poetry experts (all accomplished in publishing, critiquing, or teaching poetry) and fiction experts (equally accomplished in their field). The novices were more than 100 college students (separate from those who had written the poems or stories).

Did these novices agree with the experts? The short answer is no, although the results differed by domain. For poetry, the correlation between the two sets of raters was just $r = .22$ (Kaufman, Baer, Cole, & Sexton, 2008). The experts' ratings of the poems were fairly consistent, with a coefficient alpha of .83. The novices' ratings were far less consistent. Because coefficient alpha increases with the size of the group, we assessed what the average interrater reliability would have been for any randomly selected set of 10 novice raters. The interrater reliability of groups of 10 novices was just .58. The coefficient alpha for the full group of novice raters was .94, but getting this level of interrater agreement required 106 raters. Note that even with the full contingent of 100+ novice raters and their high coefficient alpha interrater reliability, the correlation between expert and novice ratings was still quite low. Whatever one might claim that the novices were judging, it was not creativity in poetry as understood by experts in the field.

The results were better for the short story ratings (Kaufman, Baer, & Cole, 2009). The correlation between expert and novice ratings was .71. This indicates moderate levels of agreement, certainly not acceptable for any kind of high-stakes individual assessment, but possibly high enough for group comparisons in research. The experts had high levels of interrater reliability (coefficient alpha of .92). The mean interrater reliability of randomly selected groups of 10 novice raters was just .53, but using all 106 novice raters it reached .93. It should be noted that even the moderate level of agreement between expert and novice raters required more than 100 novices. Thus, a very large number of novice raters managed to produce creativity ratings somewhat similar to experts—good enough, perhaps, for some creativity research purposes.

Our hypothesis is that novice–expert creativity agreement varies by domain, with one possible variable being the nature of the expertise required to be accomplished. Although most domains take 10 years of deliberate practice to become a creative expert (Simonton, 1997), Simonton (2009) argued for a hierarchy within domains, with "hard sciences" at one end of the extreme (highest), "soft sciences" in the middle, and arts and humanities at the other end (lowest). Some of the variables that Simonton used in his model include the level of domain consensus. If people agree about the key components needed to produce new work, then it is likely that most experts possess this knowledge. Novices without such expertise would likely be at an even greater disadvantage for these domains than for domains in which there is a high level of disagreement even at the expert level. This concept is also in accord with Amabile's (1983) suggestion that the more esoteric or specialized the field, the more narrow the range of possible experts.

WHAT DO WE KNOW ABOUT THE JUDGMENTS OF EXPERTS AND QUASI-EXPERTS?

If the outlook for novice raters is poor, the research on quasi-experts is considerably more heartening. We now review research comparing experts and what we call quasi-experts. These quasi-experts (or gifted novices) have more experience in a domain than novices, but they also lack recognized standing as experts. If an expert artist is one who has had work displayed in galleries and museums, a quasi-expert might be an MFA candidate in Art.

Many studies in the area of aesthetics have compared expert and nonexpert responses to paintings. One such study examined aesthetic judgment of student art in experts, quasi-experts, and novices (Hekkert & van Wieringen, 1996). The focus of the paper was on aesthetic preference, not agreement. They did find that the three groups tended to agree about some types of art (figurative) but not others (abstract). Indeed, expert

and novice judges were far apart in how much they liked abstract paintings, with the quasi-experts in the middle.

There have been fewer studies that specifically examine rater agreement. Amabile (1982) compared the creativity ratings made by experts and by quasi-experts. In Amabile's first study, three groups of judges rated the creativity of a small collection of collages created by 22 girls: psychologists from Stanford, art teachers who happened to be taking a course at Stanford, and artists from the Stanford art department. The latter two groups have the kind of expertise that Amabile argued was essential ("judges who have at least some formal training and experience in the target domain;" Amabile, 1996, p. 73); the group of psychologists lacked such training but might, because of their knowledge of children (and perhaps of creativity research), be thought of as having at least some related expertise. The correlation between the ratings of the artists and the psychologists was just .44. Although 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 cannot be considered complete novices. Yet the middling correlation is a caution that quasi-experts are not a perfect solution.

In Amabile's (1982) third study, she had experts and a blended group of nonexperts evaluate children's drawings. Unfortunately, her nonexpert group was a blend of novices and quasi-experts (psychology graduate students, school teachers, and undergraduates) and all statistics were based on all nonexperts. The nonexperts, like the experts, showed fairly strong agreement; the two groups were correlated at $r = .69$ (comparable to subsequent findings by Kaufman et al., 2009). Overall, these studies using collages made with precut pieces of construction paper tailored for students with limited artistic backgrounds thus show moderate levels of agreement between the ratings of various groups of quasi-experts and experts. In her 1982 paper, Amabile also reported three other studies of collages using quasi-experts of different kinds, such as students working on honors projects in studio art, but did not match them with experts so that ratings could be compared.

Amabile (1982) reported one study of the CAT using nonexpert judges that used poems as the creative product (Study 7). Unfortunately, these results are difficult to interpret for two reasons: (a) the type of poetry used—cinquains—is a form of poetry that is rarely used by actual poets and (b) the level of expertise of the nonexperts is not given in any detail. Cinquain is a staple of elementary school classrooms. Yet unless a poet is familiar with the work students do in elementary school, or happens to be a fan of the American poet Adelaide Crapsey (1922), such an "expert" judge is actually unlikely to have encountered very many cinquains. As such, it is unclear who might be properly called an expert in the field of cinquain poetry. (When Amabile described this study in her 1983 book, she changed the name of the poetry form from cinquain to "American Haiku" although she acknowledged in a footnote that these are not actually haiku and are best described as cinquains [p. 51].)

The background of the nonpoet cinquain judges was not described, other than saying that they were "five nonpoets who lived in Cambridge, Massachusetts" (Amabile, 1982, p. 1008) and that they were well educated. As Amabile wrote, "This high level of agreement might have arisen because the cinquain is so simple and because most educated individuals in our culture are familiar with simple poetic forms" (p. 1009). Educated nonpoets are perhaps as likely to have encountered cinquains as poets, so although Amabile found considerable agreement between her poet and nonpoet judges (.80), it is difficult to interpret this correlation. For this reason, even for a creativity researcher who wanted to use cinquains, it is not clear, based on these data, that true novices (e.g., undergraduates) would match poets' ratings. Quasi-expert judges (well-educated individuals such as the Cambridge nonpoets used by Amabile) might suffice, however. When Amabile discussed this study in her 1983 book, she did not mention the nonpoet judges.

Cheng, Wang, Liu, and Chen (2010) compared ratings by three different types of experts and quasi-experts: experienced teachers, teachers who had won writing awards in national contests, and professors of children's literature. All three groups had solid, if unspectacular, interrater reliabilities (.62, .69, and .58, respectively), with the experienced teachers reaching a significantly higher level of agreement that the professors. The combined interrater reliability for all raters was .85.

Plucker, Holden, and Neustadter (2008) cleverly studied expert and quasi-expert ratings by proxy in their analysis of agreement on movie ratings. They compared critic responses across different movie review compilation Web sites. Consistent with past CAT research, they found that professional Web sites (such as the National Society of Film Critics or Rotten Tomatoes) were highly correlated. Plucker et al. (2008) also examined user-driven Web sites, such as the International Movie Database (IMDb), and found significant and strong agreement with the professional Web sites (although the correlations were a bit lower). Although some of the IMDb users may be pure novices, many are likely quasi-experts; film professionals regularly visit these Web sites, along with many amateur critics and movie buffs. To see how genuine novices compared, Plucker, Kaufman, Temple, and Qian (2009) extended the original study and also asked 129 novices (college

students) to rate 680 films. These scores were then compared to reviews from professional and user-driven Web sites. Plucker et al. (2009) found that although novices and experts were significantly correlated, it was of medium strength (.43). The novice or quasi-expert ratings from the user-driven Web sites were strongly correlated with both genuine novices (.65) and experts (.72), perhaps suggesting some type of continuum of expertise does exist. Gerrard, Poteat, and Ironsmith (1996), Haller, Courvoisier, and Cropley (2010), and Runco, McCarthy, and Svenson (1994) also reported differences in creativity judgments between experts, quasi-experts, and novices, although neither study reported a correlation between different groups.

We have also studied this question using two kinds of creative products that are prominent in creativity research: poems and short stories. We first asked three groups of experts to rate a set of 27 short stories and 28 poems for creativity (Baer et al., 2004; Baer, Kaufman, & Riggs, 2009; Kaufman, Gentile, & Baer, 2005). One group (creative writers) had exactly relevant expertise, whereas the other two (creativity researchers and school teachers) had near-relevant expertise. The different types of experts generally agreed with each other. The psychologists and teachers agreed with each other at $r = .90$ (stories) and $r = .69$ (poems). Psychologists agreed with the creative writers at $r = .67$ (stories) and $r = .87$ (poems) and teachers showed an agreement with the creative writers at $r = .69$ (stories) and $r = .73$ (poems).

The gifted novices' evaluations largely agreed with those of the experts. For the gifted novices, the coefficient alpha reliabilities were .82 for the poems and .74 for the short stories (Kaufman et al., 2005). For the experts, the coefficient alpha reliabilities were .88 for the poems and .88 for the short stories. Novices ratings correlated with summed expert ratings at $r = .78$ for poetry and $r = .77$ for short stories. The correlations of the gifted novices' ratings with those of the three different types of experts across poetry and short stories ranged from .62 to .80. Gifted novices' ratings of poetry and short stories both correlated highest with the ratings of the expert group of writers.

CAN EXPERTISE BE TRAINED?

Dollinger and Shafran (2005) used a modified version of the CAT (a cross between the Test of Creative Thinking-Drawing Production and a CAT measure of creativity in drawing) and found the expert judges and novices produced fairly similar ratings. Twenty participants produced drawings in response to the Test of Creative Thinking-Drawing Production stimulus, and these 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 graduate training, although it was unknown if they studied creativity) also judged the 20 drawings. Before the ratings they first underwent a brief training activity that resembles the kinds of calibration training given to holistic scorers in other (non-creativity) assessments (Johnson, Penny, & Gordon, 2000). The nonexpert judges were shown drawings from an entirely different study and the ratings that these drawings had received from a panel of expert judges. This experience, it was hoped, would provide the novices with a framework for judging the drawings they would be rating for creativity that would be similar to the way that actual experts might evaluate them.

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).

In Dollinger and Shafran's (2005) study, novices produced similar ratings to experts (.87 for "quality of drawing" and .90 for "overall creative Gestalt"). Although a small study (just 20 drawings), it suggests that training of this kind might reduce the need for expert judges. Yet, of course, experts are still needed to train the novices. A deeper question is how much the training will transfer from domain to domain (or even from microdomain to microdomain). Would novices trained on products derived from one stimulus be able to also match expert agreement on products from a different stimulus? Perhaps. Would novices trained on products from one microdomain (such as the aforementioned haikus) be able to transfer to a related microdomain (sonnets)? Where is the line that distinguishes the need for new training? Would poetry training transfer to short stories? Likely not. Poetry training would be very unlikely to transfer to art, and even less likely to transfer to math. Depending on the number of distinct trainings needed, this concept may be of great theoretical interest but, in terms of hours and effort saved, little practical importance.

WHEN MIGHT NOVICES BE APPROPRIATE JUDGES OF CREATIVITY?

This review of what is known about the use of nonexperts as judges of creativity does not mean that novice

raters could never be used appropriately when assessing creativity using the CAT. We can think of three instances in which novice raters might be appropriately and validly employed:

1. When the focus is on the raters themselves and/or the goal is explicitly to collect judgments of novices, such as one might do in a study of how laypeople judge creativity in some domain. In this case, the judgments made are being used to learn about the people themselves.
2. When prior research has shown that novices' creativity ratings in the domain in question sufficiently match those of experts. For example, novices' ratings of the creativity of short stories in the study described above matched those of experts at a level of $r = .71$. If one were conducting a study in which two techniques for enhancing students' story-writing creativity were being compared and one had a large sample, 100 novice raters might be used instead of a panel of actual experts in the domain because only group comparisons were being made, and for this purpose, the .709 correlation might be deemed sufficient. One would need to note this, of course, when presenting the results. In effect, this would be similar to using a short group IQ test instead of a more extensive individual test. The shorter group and less valid test might have an acceptable correlation with the longer and better test and thus have sufficient validity for some testing purposes.
3. When the novices have been trained to rate the artifacts in a way that can be shown to yield ratings comparable to experts. The specific rules of the CAT expressly forbid training the experts. If one trains expert judges to rate the creativity of a product according to guidelines established by nonexperts in the domain (such as psychologists), then one has, in effect, removed the expertise of the judges. According to the principles of the CAT, therefore, one must not, therefore, in any way train experts to make creativity judgments, or give them rubrics to follow in making such judgments, or in any other way interfere with their unfettered assessments of an artifact's creativity.

   But a sequence of investigations comparable to Dollinger and Shafran (2005) could be carried out to establish a basic protocol for the exact circumstances under which trained novices could mimic experts. In this case, these alternate ratings could share the validity of expert ratings. Such trained novices would likely not work at the highest levels of creativity. However, at the more everyday (little-c) or mini-c levels of creativity that most students or research participants would be expected to exhibit (e.g., Kaufman & Beghetto, 2009), it might be achievable. Note that the researchers doing the training have not replaced the experts' judgments with their own. A system such as this that trains novices to match the way expert ratings would not be using the CAT, but its validity would be rooted in (and attested to by) CAT ratings.

THE PROMISE OF QUASI-EXPERTS AND EXPERT-LESS TASKS

Just as there are some domains where experts and novices agree at an acceptable level, so, too, are there domains where quasi-experts agree with experts. The process would be comparable as that to certify novices; one would need to show that the creativity ratings of the quasi-experts were sufficiently similar to the ratings of experts in the domain. Early research, however, indicates that quasi-experts and experts show higher levels of agreement than do novices.

Quasi-expert assessments with expert supervision have been used in other fields. One example explores ways in which collective intelligence (e.g., Wikipedia) may be very effective. Malone, Laubacher, and Dellarocas (2009) suggested a hierarchical model is often employed, in which novices may vote, but their results are screened (or used in a consultative manner) by experts who then make the actual decisions. They reported one interesting area in which quasi-experts were as successful as experts. In 2001–2002, NASA let amateur astronomers study the photos of Mars and suggest which features were craters. When the coordinates for craters proposed by the quasi-experts were averaged, they mirrored findings by expert scientists.

Just as there are quasi-experts who are neither expert nor novice, so, too, are there domains with no obvious experts. If one asks participants to write creative captions for photographs or titles for stories, who would have the appropriate expertise to judge those captions or titles? Editors, perhaps? Or might any reasonably well-read person be an appropriate expert in such instances? Kaufman, Lee, Baer, and Lee (2007) found that psychology graduate students showed a high level of agreement, both with each other and across participants (i.e., ratings were consistent across ten different captions written by the same person).

Even in expert-less domains there is variation in creativity level. We believe that such expert-less tasks may represent good choices for researchers without access to typical experts—but with a caution. The validity of creativity ratings in a domain with no readily identifiable set of experts is necessarily more tentative than

creativity ratings in domains with clear-cut levels of expertise. Comparable to Simonton's (2009) hierarchy of domains, there may be areas even further on the low end of needed expertise: Twitter messages, practical jokes, T-shirt designs, and the like. Using this continuum, we suggest that assessments of creativity in domains that have more clear-cut levels of expertise have higher potential validity, but creativity ratings in these less clear-cut domains likely also have some validity.

Whether researchers use the CAT strictly with appropriate experts or whether they are more lenient and use trained novices, quasi-experts, or expert-less domains, we think it is incumbent upon researchers to briefly discuss these issues. In domains without clear levels of expertise, they should caution readers that the validity of the ratings cannot be perfectly assessed. In domains with clear-cut experts, researchers should either describe their judges' expertise or, if novices or quasi-experts are being used, describe any existing evidence that such judges demonstrate validity in this domain.

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, NY: Springer-Verlag.

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

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

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

Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal*, *10*, 25–31.

Baer, J. (1998). Gender differences in the effects of extrinsic motivation on creativity. *Journal of Creative Behavior*, *32*, 18–37.

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.

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

Bloom, B. S. (Ed.). (1985). *Developing talent in young people*. New York, NY: Ballantine.

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

Carson, S. (2006, April). *Creativity and mental illness*. Invitational Panel Discussion Hosted by Yale's Mind Matters Consortium, New Haven, CT.

Cattell, J. Glascock, J., & Washburn, M. F. (1918). Experiments on a possible test of aesthetic judgment of pictures. *American Journal of Psychology*, *29*, 333–336.

Chen, C., Himsel, A., Kasof, J., Greenberger, E., & Dmitreiva, J. (2006). Boundless creativity: Evidence for domain generality of individual differences in creativity. *Journal of Creative Behavior*, *40*, 179–199.

Cheng, Y.-Y., Wang, W.-C., Liu, K.-S., & Chen, Y.-L. (2010). Effects of association instruction on fourth graders' poetic creativity in Taiwan. *Creativity Research Journal*, *22*, 228–235.

Child, I. L. (1962). Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality*, *30*, 496–512.

Crapsey, A. (1922). *Verse*. New York, NY: A. A. Knopf.

Csikszentmihalyi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313–335). New York: 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.

Ericsson, K. A., Roring, R. W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, *18*, 3–56.

Gardner, H. (1993). *Creating minds*. New York, NY: Basic Books.

Gerrard, L. E., Poteat, G. M., & Ironsmith, M. (1996). Promoting children's creativity: Effects of competition, self-esteem, and immunization. *Creativity Research Journal*, *9*, 339–346.

Haller, C. S., Courvoisier, D. S., & Cropley, D. H. (2010). Correlates of creativity among visual art students. *International Journal of Creativity and Problem-Solving*, *20*, 53–71.

Hayes, J. K. (1989). Cognitive processes in creativity. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Handbook of creativity* (pp. 135–145). New York: Plenum Press.

Hekkert, P., & Van Wieringen, P. C. W. (1996). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *American Journal of Psychology*, *109*, 389–407.

Hickey, M. (2001). An application of Amabile's Consensual Assessment Technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, *49*, 234–244.

Hood, R. W. (1973). Rater originality and the interpersonal assessment of levels of originality. *Sociometry*, *36*, 80–88.

Hull, D. L., Tessner, P. D., & Diamond, A. M. (1978). Planck's principle: Do younger scientists accept new scientific ideas with greater alacrity than older scientists? *Science*, *202*, 717–723.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*, 121–138.

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., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the Consensual Assessment Technique. *Journal of Creative Behavior*, *43*, 223–233.

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., & Beghetto, R. A. (2009). Beyond big and little: The Four C Model of Creativity. *Review of General Psychology*, *13*, 1–12.

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 validity. *Thinking Skills and Creativity*, *2*, 96–106.

Kaufman, J. C., Niu, W., Sexton, J. D., & Cole, J. C. (2010). In the eye of the beholder: Differences across ethnicity and gender in evaluating creative work. *Journal of Applied Social Psychology*, *40*, 496–511.

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. New York, NY: Wiley.

Kaufman, S. B., & Kaufman, J. C. (2007). Ten years to expertise, many more to greatness: An investigation of modern writers. *Journal of Creative Behavior*, *41*, 114–124.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Lee, S., Lee, J., & Youn, C.-Y. (2005). A variation of CAT for measuring creativity in business products. *Korean Journal of Thinking and Problem Solving*, *15*, 143–153.

Levin, S. G., Stephan, P. E., & Walker, M. B. (1995). Planck's principle revisited—A note. *Social Studies of Science*, *25*, 35–55.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2009). *Harnessing crowds: Mapping the genome of collective intelligence* (MIT Sloan Research Paper No. 4732–09). Retrieved from http://ssrn.com/abstract=1381502

Mayer, R. E. (1999). Fifty years of creativity research. In R. J. Sternberg (Ed.), *Handbook of human creativity* (pp. 449–460). New York, NY: Cambridge University Press.

Niu, W., & Sternberg, R. J. (2001). Cultural influence of artistic creativity and its evaluation. *International Journal of Psychology*, *36*, 225–241.

Plucker, J. A., Holden, J., & Neustadter, D. (2008). The criterion problem and creativity in film: Psychometric characteristics of various measures. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 190–196.

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way?. *Psychology and Marketing*, *26*, 470–478.

Runco, M. A., McCarthy, K. A., & Svenson, E. (1994). Judgments of the creativity of artwork from students and professional artists. *Journal of Psychology*, *128*, 23–31.

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

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*, 66–89.

Simonton, D. K. (2000). Creative development as acquired expertise: Theoretical issues and an empirical test. *Developmental Review*, *20*, 283–318.

Simonton, D. K. (2009). Varieties of (scientific) creativity: A hierarchical model of disposition, development, and achievement. *Perspectives on Psychological Science*, *4*, 441–452.

Sternberg, R. J. (1999). A propulsion model of types of creative contributions. *Review of General Psychology*, *3*, 83–100.

Sternberg, R. J., Kaufman, J. C., & Pretz, J. E. (2002). *The creativity conundrum: A Propulsion Model of kinds of creative contributions*. Philadelphia, PA: Psychology Press.