Psychology of Aesthetics, Creativity, and the Arts
2013, Vol. 7, No. 4, 332–340

1931-3896/13/$12.00 DOI: 10.1037/a0034809

Furious Activity vs. Understanding: How Much Expertise Is Needed to Evaluate Creative Work?

James C. Kaufman
University of Connecticut

John Baer
Rider University

David H. Cropley
University of South Australia

Roni Reiter-Palmon and Sarah Sinnett
University of Nebraska at Omaha

What is the role of expertise in evaluating creative products? Novices and experts do not assess creativity similarly, indicating domain-specific knowledge's role in judging creativity. We describe two studies that examined how "quasi-experts" (people who have more experience in a domain than novices but also lack recognized standing as experts) compared with novices and experts in rating creative work. In Study 1, we compared different types of quasi-experts with novices and experts in rating short stories. In Study 2, we compared experts, quasi-experts, and novices in evaluating an engineering product (a mousetrap design). Quasi-experts (regardless of type) seemed to be appropriate raters for short stories, yet results were mixed for the engineer quasi-experts. Some domains may require more expertise than others to properly evaluate creative work.

Keywords: creativity, expertise, creative domains, assessment

> Furious activity is no substitute for understanding.
>
> —H. H. Williams

Popular and critical tastes in the arts occasionally align (such as in the hit movie *Avatar*; Cameron, 2009), but more often are in opposition. For example, actors who have won the People's Choice Awards include Adam Sandler, Kate Hudson, Vince Vaughn, and other stars unlikely to impress film critics. In contrast, past acting winners of the National Board of Review Awards include Lesley Manville, Jackie Weaver, Emile Hirsch, and other acclaimed but comparatively obscure thespians.

One reason for this discrepancy is the different levels of expertise needed for evaluating creative products. For example, the movie critics who vote on the National Board of Review Awards (and the industry members who vote on the Academy Awards) have a much higher level of expertise than the average layperson. Both the critics and the professionals have spent years either making or evaluating movies, comparable to the 10 years of deliberate practice needed to make a substantial contribution to a field (Ericsson, Roring, & Nandagopal, 2007). Most laypeople, however, are novices. Somewhere in the middle are "quasi-experts." We define quasi-experts as individuals with more experience in a given domain than novices, but who do not have a recognized standing as experts. In the domain of movies, they might include amateur movie buffs or people in the film industry who might be more removed from actual filmmaking (i.e., gaffers).

Plucker, Holden, and Neustadter (2008) compared movie reviews from professional critics (experts) to scores on user-driven websites such as the International Movie Database (IMDb.com); these raters could be considered quasi-experts. Plucker, Kaufman, Temple, and Qian (2009) extended the original study with a group of college students with no particular experience with film. The critics showed the strongest reliability, followed by IMDb users, and then the students. The IMDb users were highly correlated with both students (.65) and critics (.72). However, the correlation between critics and students was notably lower (.43).

The question of how expert, quasi-expert, and novice aesthetic judgments are related goes far beyond film, of course. There are also important implications for creativity. For example, extensive work has looked at differences in actual creative performance and problem solving. Some research has examined the actual creative process (e.g., Voss, Wolfe, Lawrence, & Engle, 1991), whereas other studies have investigated the relationship between expertise

James C. Kaufman, Neag School of Education, University of Connecticut; John Baer, Department of Teacher Education, Rider University; David H. Cropley, School of Engineering, University of South Australia, Adelaide, SA, Australia; Roni Reiter-Palmon and Sarah Sinnett, Department of Psychology, University of Nebraska at Omaha.

We thank Alexander S. McKay, Kristen Ramos, Paul Silvia, Dean Keith Simonton, and Arielle White for their assistance and suggestions with earlier drafts of the manuscript.

Correspondence concerning this article should be addressed to James C. Kaufman, Professor of Educational Psychology, University of Connecticut, Neag School of Education, 2131 Hillside Road, Storrs, CT 06269-3007. E-mail: james.kaufman@uconn.edu

and rigidity in problem solving (e.g., Bilalić, McLeod, & Gobet, 2008; Schooler & Melcher, 1995).

The impact of expertise on a person's reaction to a creative product has been less frequently examined. In one classic paper, Hull, Tessner, and Diamond (1978) studied the ages of both early accepters and continued rejecters of Darwin's theory of evolution as proposed in *The Origin of Species* (1859). Based on these results, they proposed Planck's principle, which argues that younger scientists are more likely to accept new ideas than older scientists (a subsequent study, however, found contradictory results; see Levin, Stephan, & Walker, 1995).

Another way of approaching this issue is to look at creativity assessment. One common way to measure creativity is to ask raters to evaluate actual creative work. Initially called "aesthetic judgment" and emphasizing the arts, this work initially started nearly 100 years ago (Cattell, Glascock, & Washburn, 1918; Child, 1962). More recently, Amabile (1982, 1996) established specific guidelines for using raters to evaluate creativity, which she dubbed the consensual assessment technique (CAT). The CAT is typically used for artistic efforts, including collages (Amabile, 1996; Hennessey, Kim, Guomin, & Weiwei, 2008), short stories and poetry (Baer, Kaufman, & Gentile, 2004; Baer, Kaufman, & Riggs, 2009; Kaufman, Gentile, & Baer, 2005), photo captions (Kaufman, Lee, Baer, & Lee, 2007), photographic essays (Dollinger, 2007), designs (Haller, Courvoisier, & Cropley, 2010), dramatic performances (Myford, 1989), and music compositions (Hickey, 2001). It is less common to apply the CAT to nonartistic work, although there have been several studies demonstrating its effectiveness in such domains as writing mathematical problems and equations (Baer, 1994), inventions (Cropley & Kaufman, 2012), responding to science-based questions about an animal's habitat (Kaufman, Evans, & Baer, 2010), or solving everyday problems (Reiter-Palmon, Mumford, Boes, & Runco, 1997). Conceivably, the CAT could be used on any type of creative product.

According to the CAT, creative work should be assessed by experts. The question of what type of expert should be used is its own debatable topic. Conceivable experts for judging high school poetry could include professional poets, creative writing teachers, literary journal editors, and perhaps even experienced English teachers or creativity researchers. Most research has found that different types of experts have solid-to-strong interrater reliability (Amabile, 1996; Baer et al., 2004; Cheng, Wang, Liu, & Chen, 2010). In addition, different expert groups (i.e., teachers and writers) tend to agree with each other, with correlations typically higher than $r = .40$ and often above $r = .70$ (Amabile, 1996; Baer et al., 2009; Getzels & Csikszentmihalyi, 1976).

The relationship between expert and novice ratings of creative work is less convergent. Lee, Lee, and Young (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 among the novices. They also calculated that the variance due to raters was much lower for the experts, also indicating a higher level of agreement. Finally, Lee et al. (2005) found that product-based variance—differences between the ratings given to different products—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.

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.

A series of studies investigated this question by domain. Kaufman, Niu, Sexton, and Cole (2010) collected both poetry and short stories by more than 200 college students. Both samples of work were then rated by two different groups of judges: experts and novices. There were poetry experts (all accomplished in publishing, critiquing, or teaching poetry) who rated the poems and fiction experts (equally accomplished in their field) who rated the short stories. The novices were more than 100 college students (separate from those who had written the poems or stories).

The novices were not found to be comparable to the experts, although the extent of divergence related to the 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 α of .83. The novices' ratings were far less consistent. Because coefficient a increases with the size of the group, the authors 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 α for the full group of novice raters was .94, but getting this level of interrater agreement required 106 raters. Even with the full contingent of more than 100 novice raters and their high α interrater reliability, however, the correlation between expert and novice ratings was still low. Whatever one might argue was the basis of the novices' judgments, 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 α 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 (.71) 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, but limiting.

Why might novices and experts disagree on creative work? Much of the work on this question has focused on visual art and emphasizes different emotional and intellectual responses. Leder, Gerger, Dressler, and Schabmann (2012) found that experts and novices have different emotional responses to art, and emotion played a bigger role for novices in their appraisal of art. Silvia (2006) compared novices to quasi-experts (people with some background and training in the arts). He found that quasi-experts were better able to understand complex pictures and, thus, also found them more interesting (Millis, 2001, found comparable results).

Locher, Smith, and Smith (2001) compared how people with and without art training viewed classic paintings. The experts (with training) saw the paintings as being more complex, varied, interesting, and distinct.

Other reasons may be found in psychophysiological work, such as Müller, Höfel, Brattico, and Jacobsen's (2010) study of novices and expert approaches to music. They physiologically measured how experts and novices responded to music. Müller et al. (2010) found that experts and novices showed different brain responses to different types of chord progressions. Their results suggested that experts are better at perceiving music and can more easily switch between types of listening modes.

Most creativity researchers would prefer to use novices instead of experts if possible because of very practical reasons (Kaufman, 2009). Experts are hard to find, may expect payment, and may be less likely to agree to rate a large number of items. Yet the research suggests that simply using novices is not an appropriate substitute. One possible compromise might be to use quasi-experts as raters. Quasi-experts, as discussed earlier, have more experience in a domain than novices, but also lack recognized standing as experts. If an expert poet has published in many literary journals and given multiple readings, a quasi-expert might be an MFA candidate in creative writing emphasizing poetry. The studies on the use of quasi-experts are encouraging to those who want to use expert raters but are encumbered by the practical issues involved (such as expense and time).

Hekkert and van Wieringen (1996) looked at expert, quasi-expert, and novice aesthetic preference. They found 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. Amabile (1996) reports a series of studies looking at experts, quasi-experts, and novices; although there is often not enough information about the specific level of expertise of the raters to make sweeping conclusions (see Kaufman & Baer, 2012, for a detailed discussion), quasi-experts generally show agreement with experts. Kaufman et al. (2005) used gifted novices and experts to rate stories and poetry; they found strong (if lower) interrater reliabilities for the gifted novices, with correlations of $r = .78$ for poetry and $r = .77$ for short stories.

The purpose of the first study reported in this paper was to investigate how quasi-expert judgments of creative work compared with both novice and expert judgments. We predict that quasi-experts, regardless of their nature of expertise, will give more reliable ratings than novices (H_1). Further, these ratings will be more highly correlated with expert ratings than novice ratings (H_2). Finally, expert ratings will continue to show the highest levels of reliability (H_3).

The second study will then see how these findings transfer to another domain.

Study 1

Method

Participants. Four groups of quasi-experts were recruited for this study. The first group, creativity students, consisted of 12 advanced undergraduate and graduate students (eight women and four men) actively involved in creativity research. All 12 students had had training in creativity research, had participated in conducting experiments in creativity, and in most cases (nine of 12) had presented their research at professional conferences. Professional creativity researchers have been used as experts in past work (Baer et al., 2004), and their ratings strongly correlated with domain experts (Baer et al., 2009).

There were three other groups of quasi-experts whose quasi-expertise was based on their connection to the field of teaching. One of these groups consisted of 10 elementary education majors (all were sophomores or juniors, and all were women) who were preparing to be elementary teachers. Although not English majors, this group had at least some experience in reading and grading student papers and in the study of creativity in their educational psychology coursework. A second group of teacher quasi-experts with a somewhat greater degree of expertise included 10 junior and senior secondary education/English double majors (one man, nine women) who were preparing to be English teachers. The third group of teacher quasi-experts consisted of nine currently employed English teachers with at least two years of experience teaching English (six women, three men). Each of these quasi-expert raters received a small honorarium for their participation.

The material was taken from a past study (Kaufman, Niu, et al., 2010) and consisted of 205 short stories written by college students. The sample that generated the stories included 54 men and 151 women, with a mean age of 24.2 years ($SD = 8.73$ years).

Expert and novice ratings were taken from a past study (Kaufman et al., 2009). The expert raters from the past study were 10 professional writers (seven women, three men). Five had master of fine arts (MFA) degrees in creative writing and three others had PhDs in English; all 10 had been published. Consistent with Amabile's (1996) consensual assessment technique, they never met or discussed their ratings in any way.

Novice raters consisted of 106 college students from a California public university who participated in the study for course credit. The novice 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.

Procedure. All raters received the same instructions for rating the short stories, as follows, modeled after past studies (Kaufman, Plucker, & Baer, 2008):

1. There were two prompts: "Execution" and "2305." Students could choose either one and were asked to write a story with one of the two prompts as its title.
2. Please rate each story for creativity on a scale of 1–10, with 1 being *least creative*, 10 the *most creative*.
3. Please compare the stories to one another, not to some other standard. These are not stories written by students who identify themselves as writers. They were subjects in a psychology experiment.
4. Please try to use the full scale.
5. There is no need to explain or justify your ratings. Use your expert sense to judge each story for its creativity (not for spelling, punctuation, etc.—just creativity).

Data analysis. The effectiveness of different rater groups was evaluated using two methodologies. The first method evaluated the reliability of the ratings for each of the groups. Interrater reliabilities were evaluated using Cronbach's α. In addition, as the number of novice raters was exceedingly large (which would increase their interrater reliability), the reliabilities of random smaller samples were evaluated. Specifically, using a random number generator, 10 novices were selected at random (equal to the number of expert raters), and the interrater reliability was calculated on this subsample. The random sampling was repeated 100 times.

Scores from the three groups of raters were correlated to determine whether the rank order of the rated targets was similar across groups of raters, since correlations evaluate rank order and cannot be used to compare mean similarities or differences. High correlations would indicate that the raters across the different groups rated the targets (stories) in a similar fashion, without evaluating the actual ratings, only the rank order of those ratings. The same random samples were used to calculate correlations. To determine the average across the random samples, Pearson correlation coefficients (*r*) were transformed using Fisher's *Z*, averaged, and then the average was transformed back to *r*.

Results and Discussion

Reliabilities were high across all rater groups when using the full group (experts .93; quasi-experts .97; and novices .99). When the 41 quasi-experts were broken down into the four specific subgroups, reliabilities were somewhat lower (likely due to the smaller sample sizes of 12, 10, 10, and 9) but were still high (.86 to .92). When evaluating the reliabilities for the novices using random samples of 10 novices[1], however, a very different picture emerged. The reliabilities ranged from .35 to .75, with an average reliability across the 100 samples of .53. These results indicate that a very large sample of novices may be as reliable as experts; however, when smaller sample sizes are used (as is the case in most creativity studies), novices are not reliable. In fact, novices perform fairly poorly. The highest interrater reliability of the novice samples that was obtained was .75, but this occurred in just one of the 100 samples.

The correlations between the rater groups are presented in Table 1. All the correlations are high, indicating some degree of overlap in the rank order of ratings. As can be seen, the correlation between experts and full group of quasi-experts was high (r = .89), indicating that the two groups rate in a similar fashion. The lowest correlation was seen between the experts and the 106 novices (r = .72). Further, when the specific subgroups of quasi-experts were examined, the relationships between expert ratings and quasi-expert ratings remained strong, especially for English teachers and creativity students. It is notable that these relationships continued to be high, despite the fact that the specific groups were smaller (about 10–12 quasi-experts, comparable to the number of experts in this study and similar to the number of judges frequently used in CAT-based research studies).

Table 1
Study-1 Correlations Between Rater Groups

Scale	1	2	3	4	5	6	
Experts	—						
Quasi-experts overall	.89	—					
Quasi-expert English teachers	.86	.95	—				
Quasi-expert English majors	.77	.93	.84	—			
Quasi-expert education majors		.80	.94	.87	.83	—	
Quasi-expert creativity students		.91	.95	.87	.83	.84	—
Novices overall		.72	.83	.81	.77	.81	.73

The results obtained using the 100 random samples of 10 novices provide a quite different picture. The lowest correlation between experts and novices was .22 and the highest was .76, with an average of .56. The correlation between quasi-experts and novices ranged from .32 to .85, with an average of .66. These results indicate that although experts and quasi-experts have a high degree of agreement regarding the creativity of the stories as indicated by the correlations, novices' judgments are not similar to those of experts. The overall correlation (using the full novice sample) was still lower than quasi-experts. When smaller samples were used, the correlations between experts and novices were even lower. Taken together, this study suggests that novices are not appropriate substitutes for experts in providing creativity ratings (consistent with Kaufman et al., 2009). However, quasi-experts seem to provide similar ratings to those of experts, both in terms of reliability and rank order, and therefore can be used as substitutes for experts.

Study 2

Although quasi-experts seem to be suitable substitutes for experts in the domain of creative writing (Study 1), there is an extensive literature on how creativity differs across domains (see Kaufman & Baer, 2005). It is important to explore whether these findings would transfer to other domains. In particular, one key question is whether these results would vary in a domain that requires a different level of expertise to become accomplished.

Although most domains take 10 years of deliberate practice to become a creative expert (Simonton, 1997), Simonton (2004, 2009) argues for a hierarchy within domains based on seven dimensions ranging from citation concentration to theories-to-laws ratio. One of these key dimensions, he emphasizes, is peer-evaluation consensus. A high-consensus field would indicate that the body of knowledge within the domain is well-defined and that most experts within this domain would have such knowledge. One example given by Simonton (in press) is that, even though Albert Einstein was considered an outsider, his achievements were recognized by his peers. Simonton's hierarchy focuses on science, with the following sciences listed from highest place in the hierarchy to lowest: physics, chemistry, biology, psychology, and sociology. Extrapolating from this, arts and humanities would be at a very low level (Simonton, 2009). This concept is also in accord with Amabile's (1982) suggestion that the more esoteric or specialized the field, the more narrow the range of possible experts.

Simonton (2009) placed the sciences higher than the arts and "hard" sciences (e.g., physics, chemistry) higher than "soft sciences (e.g., sociology, psychology). Klavans and Boyack (2009), in their mapping of the sciences, placed both physics and chemistry as being strongly tied to engineering. Indeed, the nature of the domain of engineering includes very strong consensus about what

[1] This analysis is new and more detailed than the original analysis presented in Kaufman et al, 2009.

is part of its body of knowledge. This is controlled, and codified, in two ways in countries like the United States and Australia. First, national accreditation bodies like the Accreditation Board for Engineering and Technology (ABET) in the United States set out what knowledge is required to turn novices into (quasi) experts through the process of university education. Second, a variety of international, national and/or state-based professional bodies then certify practitioners at one or more levels of expertise. The International Council on Systems Engineering (INCOSE), for example, defines three levels of professional systems engineer: *supervised practitioner*, *practitioner,* and *expert*. Entry into any one of these requires a formal assessment of the candidate's *fluency* (his or her level of familiarity with the domain's body of knowledge), domain-relevant education, and practical experience.

This process may be contrasted with the way in which a creative writer achieves expertise. A creative writer may become proficient by formal schooling, informal mentorship, repeated practice, and the metric for being a professional writer (namely, getting published) may vary widely (Kaufman, 2002; Kaufman & Beghetto, 2009). That is not to say that one is any more or less expert than the other; only that the manner in which the body of knowledge in each domain is defined is very different, and the means by which a person moves from one level (novice, quasi-expert and expert) to another is very distinctive.

The manner in which creativity is understood across the spectrum of domains may also be very different. In domains like engineering, the greater level of consensus and codification of the body of knowledge imposes constraints on creativity that may not be present in domains like creative writing. Cropley and Cropley (2005, 2008, 2010) sought to address this characteristic of creativity, relative to domains, by focusing on *functional* creativity. They argued that, in the practical world of engineered products, processes, systems, and services, the most important aspect of an artifact that excites admiration in the beholder is *not* novelty, but the product's ability to meet customer needs, that is, its *effectiveness*. An automobile, for example, must transport people quickly, economically and comfortably over long distances. If it fails to satisfy requirements like these, then it lacks effectiveness and thus cannot be regarded as creative, no matter how novel it is. This reflects the constraints imposed by a highly codified domain at the upper end of Simonton's spectrum. Creativity in engineering is permitted, and valued, but only in certain ways, and at certain times.

The purpose of Study 2, then, was to explore the nature of judgments of creative work by quasi-experts, in the domain of engineering, in comparison to novices and experts. Will the findings from domains such as creative writing transfer to a domain such as engineering where the boundaries between levels are thought to be greater and more sharply delineated? In Study 2 we predict that quasi-experts will give more reliable rating than novices (H_1) and that these rating will be more highly correlated with experts than with novices (H_2). We further predict that expert ratings will show the highest levels of reliability (H_3).

Method

Participants. Three groups were recruited for this study: a small group ($N = 15$, all male) of Australian professional engineers, serving as experts; a larger group ($N = 31$, 4 females) of Australian first-year undergraduate engineering students, serving as quasi-experts; and a large group ($N = 274$, 34 males, 226 females, 14 did not provide gender information) of American students taking a psychology class at a California public university, who served as novices.

Two key characteristics of judges emerge as the criteria by which experts, quasi-experts, and novices are differentiated from each other in this study. Amabile (1996) notes the importance of "experience with the domain in question" (p. 41), and Stein (1974) highlights the acceptance of the product by an organized group within a domain. For the purposes of Study 2, novices were therefore selected on the basis that they possessed neither domain-relevant experience (in this case engineering) nor were they members of an organized, recognizable, domain-relevant group (either professional engineers or engineering students).

Membership of the expert group was restricted to participants who had a minimum of a bachelor of engineering (BE) degree and at least 10 years of experience as a professional engineer (satisfying the requirement for experience within the domain) and who held a professional membership (CPEng) with Engineers Australia (satisfying the requirement for membership of a domain-relevant group).

Membership of the quasi-expert group was restricted to participants who were enrolled in the first year of a BE degree. Although the quasi-experts also satisfy the two requirements: domain experience and membership of a domain-relevant group, they differ from the experts primarily by virtue of the *degree* of their experience. It seems reasonable to assume that first-year engineering students possess greater domain experience compared with undergraduate psychology students, but less than degree-qualified, professional engineers. The question of whether this is sufficient to distinguish quasi-experts from novices is, of course, central to the investigation.

Procedure. Participants were directed to a website where the measures were hosted online. Participants were presented, sequentially, with an image of one of five different mousetraps of varying designs. Images of the mousetraps were selected from Google image search to represent a diverse range of possible mousetraps. Participants were asked to rate each of the different mousetraps using the following five constructs:

> Overall creativity—the degree to which the mousetrap is creative, using your own subjective definition of creativity.
>
> Relevance/effectiveness—the degree to which the mousetrap fulfills the function for which it was designed.
>
> Novelty—the degree to which the mousetrap is original and surprising.
>
> Elegance—the degree to which the mousetrap is well-made, complete and Pleasing to the eye.
>
> Genesis—the degree to which the mousetrap opens up new perspectives and the problem.
>
> Each item was rated using a 5-point Likert-type scale (ranging from *very low* through *medium* to *very high*) to indicate the degree to which the item applies to the given mousetrap.

Data analysis. In a similar fashion to Study 1, the effectiveness of the different rater groups (novice, quasi-expert and expert) was evaluated by assessing both the reliability of the ratings for each group, and the correlations between groups. Also, consistent with Study 1, the reliabilities of 100 random subgroups of 10 novices were selected to filter out the effects of group size on reliability. Table 2 shows the results for reliability. The table shows values both for overall creativity (comparable to the ratings of creative writing made in Study 1) and values for the four dimensions of a creative product, defined by Cropley and Cropley (2010) in their model of functional creativity. Table 3 presents the correlations between pairs of groups (expert + quasi-expert; quasi-expert + novice; expert + novice). In both Tables 2 and 3, results are shown for the full group of novices, as well as for random samples. Similar to Study 1, to obtain the average correlation between the novice random samples and the other two groups (experts and quasi-experts), these correlations were first transformed using Fisher's Z, then averaged, and the value was then transformed into a correlation which was used in Table 3.

In relation to overall creativity, reliabilities for the three groups showed the same pattern as Study 1. Novices had the highest level of reliability (.98), followed closely by quasi-experts (.93). Experts had an acceptable level of reliability (.86), but lower than that of the other two groups, likely as a result of the low number of raters relative to the other two groups. Using random samples of novices, the results shifted, and the average reliability was considerably lower (.66) and now well below that of the other groups. Of particular note is the range of values of average reliability in the samples of novices. The lowest average reliability for overall creativity for the novices was close to zero (.06).

The correlations for overall creativity between experts and quasi-experts were moderate ($r = .52$), indicating that for this task, experts and quasi-experts differ to some extent in their ratings. The correlations indicated that novices (when using the full sample) were very similar to quasi-experts ($r = .96$). When random samples of novices were used, the range of correlations with quasi-experts was large ($r = .27$ to $r = .99$); however, the average correlation between samples of novices and quasi-experts remained high ($r = .85$). Correlations between experts and novices were more discrepant. The correlation between novices and experts, using the full sample of novices, was low ($r = .29$) indicating little overlap between expert and novice ratings, further, using the random samples of novices, some of the correlations were negative!

When evaluating the reliabilities for the other rating scales (see Table 2), the full sample of novices performed reliably (>.94); indeed, they typically outperformed experts and quasi-experts by a small margin. Overall, all three groups showed good interrater reliability as measured by Cronbach's α. However, when random samples for novices were analyzed, as in Study 1, a different picture emerged. As with the reliability for overall creativity, the minimum value obtained for reliability for each of the scales was close to zero. Although the maximum reliability was adequate, this level of reliability was seen in only a fraction of random samples. The average reliabilities based on all 100 samples ranged from .48 to .63; most would not be considered adequate for ratings of creativity. In contrast to the reliability measures for overall creativity, a less stable picture emerges for the comparison of reliabilities between experts and quasi-experts. In two cases (elegance, novelty) the reliability of experts exceeded that of quasi-experts, whereas in the remaining two cases (genesis, relevance/effectiveness) the reliability of quasi-experts exceeded that of experts.

The correlation data for the other rating scales (see Table 3) indicate, in general, that quasi-experts provided ratings that were more similar to those of novices than those of experts. The relationships between expert ratings and those of quasi-experts and novices were, however, somewhat varied depending on the specific scale that was used. For each rating scale, experts tended to show a higher correlation with quasi-experts than with novices (concentrating on the full samples), however the magnitudes of these correlations were highly variable, ranging from .78 (experts/quasi-experts: relevance/effectiveness) and .79 (experts/novices: relevance/effectiveness) to .02 and −.16, respectively for genesis. When the random samples of novices were used to evaluate correlations, the range of correlations was very large, with all scales having negative correlations between novices and experts as well as quasi-experts at the minimum values. Based on these random samples of novices, all correlations between experts and quasi-experts were higher than between experts and novices. Again, these results indicated, in general, stronger similarities in ratings between the quasi-experts and novices, likely due to the large number of novices used as raters in this study; however there were some noteworthy exceptions that will be discussed below.

Discussion

The results for Study 2 support the argument that the differences between novices, quasi-experts, and experts are larger and sharper for a more highly structured and codified domain (i.e., *hard sciences* in Simonton's hierarchy; Simonton, 2004, 2009). The net effect of this is that, in contrast to Study 1, quasi-experts, as defined in this study, were poor substitutes for experts for the purpose of evaluating overall creativity. The evaluations of overall

Table 2
Study 2 Reliabilities

Scale	Experts	Quasi-experts	Novices	Novices—random sample		
			Total sample	Minimum	Maximum	Average
Overall creativity	.86	.93	.98	.06	.88	.66
Elegance	.96	.79	.97	.15	.89	.63
Genesis	.85	.87	.96	.06	.78	.53
Novelty	.93	.90	.94	.06	.86	.48
Relevance/effectiveness	.83	.97	.97	.06	.89	.59

Table 3
Study-2 Correlations Between Rater Groups

Scale	Experts/quasi-experts	Quasi-experts/novices				Experts/novices			
		Full sample	Minimum	Maximum	Average	Full sample	Minimum	Maximum	Average
Overall Creativity	.52	.96	.27	.99	.85	.29	−.45	.89	.35
Elegance	.60	.96	−.38	.99	.86	.38	−.30	.70	.42
Genesis	.02	.92	−.48	.99	.71	−.16	−.78	.99	−.05
Novelty	.71	.75	−.80	.98	.55	.08	−.90	.96	.15
Relevance/Effectiveness	.78	.88	−.79	.99	.71	.79	−.82	.99	.65

creativity of the quasi-experts in Study 2 showed a much higher correlation with the evaluations of novices (.85) than with experts (.52). It should be noted that there were substantial differences in the gender ratios of the different groups in the two studies. Given that some gender differences in creativity do exist (Baer & Kaufman, 2008), future studies should aim for more equal numbers when possible.

A somewhat more varied picture emerges for the evaluation of the four characteristics of creativity. Two of these characteristics (elegance and genesis) showed the same pattern as for overall creativity, namely, that quasi-experts were a poor substitute for experts. For genesis, in particular, the correlation between experts and both quasi-experts (.02) and novices (-.05) was, in effect, nonexistent. Conversely, one characteristic (novelty) showed a different pattern when compared with overall creativity. In this case, quasi-expert ratings correlated quite highly with those of experts (.71) and did so more strongly than with novices (.55). Novelty also showed high reliability (.90). Finally, the correlations for relevance/effectiveness showed a third variation. In this case, ratings for all three groups were moderately strong and broadly comparable (.78, .71, .65). This also suggests that quasi-experts could function as substitutes for experts (noting also a reliability of .97 for quasi-experts for this characteristic). Indeed, it suggests that of all the measured criteria, relevance/effectiveness is the only one where an argument can be made that novices are capable of substituting for experts, to a limited degree. These results support previous findings on the evaluation of overall creativity, are consistent with hypotheses based on Simonton's (2004, 2009) hierarchy of domains, and provide some interesting insights into the role that expertise might play in relation to evaluating more differentiated, and domain-specific, criteria of creativity. There is widespread agreement (e.g., Kaufman, 2009) that the two key components of creativity are novelty and appropriateness to the task at hand (likely assessed by the relevance/effectiveness factor). In Study 2, it emerges that, unlike the overall construct *creativity*, quasi-experts may be much closer to experts in their ability to recognize what is new, original, and surprising. This finding may suggest that at least some domain knowledge (first-year engineering) is sufficient, perhaps acting in concert with a general predisposition to the domain, to form accurate judgments of novelty. Even more interesting is the fact that very little domain knowledge may be required to form a reasonable judgment of appropriateness—even those raters with no domain-specific knowledge may recognize if an artifact will do what it is supposed to do. By contrast, the criteria elegance and genesis are more sophisticated and nuanced characteristics of creativity and the results of Study 2 suggest that quasi-experts and novices are not able form judgments of these that are comparable to experts.

Overall Discussion

In summary, whereas Study 1 supports the possible use of quasi-experts as substitutes for experts in relation to the evaluation of creativity in domains that require lower levels of domain-based knowledge to acquire expertise, Study 2 suggests that the larger and sharper delineations of expertise, quasi-expertise, and absence of expertise found at the higher end of the hierarchy of domains preclude the use of quasi-experts as substitutes for experts in the evaluation of creativity. However, when creativity is broken down into more highly differentiated components, i.e., novelty, relevance and effectiveness, and elegance and genesis, a more complex picture emerges. It appears that the core criteria defining creativity (novelty and appropriateness), may be more independent of the level of expertise of the observer, whereas more sophisticated (and domain-dependent) criteria, such as elegance and genesis, remain strongly linked to the level of expertise of the observer.

Future research might investigate both the transferability of the latter finding back to other domains, such as creative writing. Many studies have had experts (Amabile, 1996; Müller et al., 2010) and novices (Rawlings, Barrantes-Vidal, & Furnham, 2000; Turner & Silvia, 2006) rate nuanced aspects of artistic work, from emotional response to technical quality. Amabile (1996), for example, asked experts to rate artistic work on such dimensions as novel use of materials, variation in shapes, symmetry, representationalism, silliness, detail, and evidence of effort and planning.

Another area for future work might be to investigate the role that different levels of quasi-expertise might play. In other words, is there a specific range of knowledge that is needed to serve as a proxy for an expert rater? Does this amount of knowledge vary by domain? Some of this work has been done on how people evaluate their own work. Silvia (2008) asked people to pick their best responses to a divergent thinking task and then examined if the chosen responses were the same as the responses chosen by outside raters. Silvia found that people were able to discern their more creative responses reasonably well; in addition, people more open to experience were more likely to choose accurately. At the Big-C end of the spectrum, Kozbelt's (2007) analysis of Beethoven's self-critiques found that the great composer was a reasonably accurate rater of his own work.

Finally, it would be interesting to see how differing familiarity with creativity scholarship impacts ratings. Although one of the

core tenets of the consensual assessment technique has been to not provide training (Kaufman et al., 2008), this question should still be explored. Dollinger and Shafran (2005) trained novice judges[2] on aesthetic judgment by showing them drawings from an entirely different study and the ratings that these drawings had received from a panel of expert judges. They then compared the trained novice ratings and expert ratings, and found the novices had strong reliability and generally agreed with the experts. The question of whether such training could be done online (thereby using less resources) and how much training is needed to increase novice or quasi-expert judgments has yet to be answered.

Conclusion

On a practical level, expense of getting expert raters to look at creative work has largely limited the consensual assessment technique to research use. Finding a happy medium might encourage more work with this technique, which allows a domain-specific perspective on creativity. On a theoretical level, exploring the question of how differing levels of expertise leads to similar or divergent perceptions of creative work can yield insight into many aspects of cognition.

This study compared novice, quasi-expert, and expert ratings of creativity in two domains, creative writing (short stories) and engineering (product design). Novices, consistent with extensive past work (e.g., Kaufman & Baer, 2012) only showed acceptable levels of reliability and expert agreement when used in excessively large numbers. When the level of analysis was lowered to groups of 10, reliability and agreement was drastically reduced. The utility of quasi-experts varied by domain. For creative writing, quasi-experts showed strong reliability and expert agreement; in engineering, the results were inconsistent and generally did not support the use of quasi-experts as raters.

[2] Possibly quasi-expert using our definitions.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43,* 997–1013. doi:10.1037/0022-3514.43.5.997

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

Baer, J. (1994). Divergent thinking is not a general trait: A multidomain training experiment. *Creativity Research Journal, 7,* 35–46. doi:10.1080/10400419409534507

Baer, J., & Kaufman, J. C. (2008). Gender differences in creativity. *The Journal of Creative Behavior, 42,* 75–105. doi:10.1002/j.2162-6057.2008.tb01289.x

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. doi:10.1207/s15326934crj1601_11

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

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts: Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology, 56,* 73–102. doi:10.1016/j.cogpsych.2007.02.001

Cameron, J. (director). (2009). *Avatar* [motion picture]. United States: Twentieth Century Fox Film Corporation.

Cattell, J., Glascock, J., & Washburn, M. F. (1918). Experiments on a possible test of aesthetic judgment of pictures. *The American Journal of Psychology, 29,* 333–336. doi:10.2307/1414125

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. doi:10.1080/10400419.2010.481542

Child, I. L. (1962). Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality, 30,* 496–512. doi:10.1111/j.1467-6494.1962.tb02319.x

Cropley, D. H., & Cropley, A. J. (2005). Engineering creativity: A systems concept of functional creativity. In J. C. Kaufman, & J. Baer (Eds.), *Faces of the muse: How people think, work and act creatively in diverse domains* (pp. 169–185). Hillsdale, NJ: Erlbaum.

Cropley, D. H., & Cropley, A. J. (2008). Elements of a universal aesthetic of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 2,* 155–161. doi:10.1037/1931-3896.2.3.155

Cropley, D. H., & Cropley, A. J. (2010), Functional creativity: Products and the generation of effective novelty. In J. C. Kaufman, & R. J. Sternberg (Eds.), *Cambridge handbook of creativity* (pp. 301–318). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511763205.019

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: empirical validation of the Creative Solution Diagnosis Scale (CSDS). *Journal of Creative Behavior, 46,* 119–137.

Darwin, C. (1859). *On the origins of species by means of natural selection*. London: Murray.

Dollinger, S. J. (2007). Creativity and conservatism. *Personality and Individual Differences, 43,* 1025–1035. doi:10.1016/j.paid.2007.02.023

Dollinger, S. J., & Shafran, M. (2005). Note on the consensual assessment technique in creativity research. *Perceptual and Motor Skills, 100,* 592–598. doi:10.2466/PMS.100.3.592-598

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. doi:10.1080/13598130701350593

Getzels, J., & Csikszentmihalyi, M. (1976). *The creative vision: A longitudinal study of problem-finding in art*. New York, NY: Wiley.

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

Hekkert, P., & Van Wieringen, P. C. W. (1996). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *The American Journal of Psychology, 109,* 389–407. doi:10.2307/1423013

Hennessey, B. A., Kim, G., Guomin, Z., & Weiwei, S. (2008). A multicultural application of the consensual assessment technique. *The International Journal of Creativity & Problem Solving, 18,* 87–100.

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. doi:10.2307/3345709

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. doi:10.1126/science.202.4369.717

Kaufman, J. C. (2002). Dissecting the golden goose: Components of studying creative writers. *Creativity Research Journal, 14,* 27–40.

Kaufman, J. C. (2009). *Creativity 101*. New York, NY: Springer.

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). Mahwah, NJ: Erlbaum.

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal, 24,* 83–91. doi:10.1080/10400419.2012.649237

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior, 43,* 223–233. doi:10.1002/j.2162-6057.2009.tb01316.x

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. doi:10.1080/10400410802059929

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology, 13,* 1–12. doi:10.1037/a0013688

Kaufman, J. C., Evans, M. L., & Baer, J. (2010). The American idol effect: Are students good judges of their creativity across domains? *Empirical Studies of the Arts, 28,* 3–17. doi:10.2190/EM.28.1.b

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. doi:10.1177/001698620504900307

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. doi:10.1016/j.tsc.2007.04.002

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. doi:10.1111/j.1559-1816.2009.00584.x

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

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60,* 455–476. doi:10.1002/asi.20991

Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music, 35,* 144–168. doi:10.1177/0305735607068892

Leder, H., Gerger, G., Dressler, S. G., & Schabmann, A. (2012). How art is appreciated. *Psychology of Aesthetics, Creativity, and the Arts, 6,* 2–10. doi:10.1037/a0026396

Lee, S., Lee, J., & Young, C.-Y. (2005). A variation of CAT for measuring creativity in business products. *Korean Journal of Thinking & 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,* 275–283.

Locher, P. J., Smith, J. K., & Smith, L. F. (2001). The influence of presentation format and viewer training in the visual arts on the perception of pictorial and aesthetic qualities of paintings. *Perception, 30,* 449–465. doi:10.1068/pp.3008

Millis, K. (2001). Making meaning brings pleasure: The influence of titles on aesthetic experience. *Emotion, 1,* 320–329. doi:10.1037/1528-3542.1.3.320

Müller, M., Höfel, L., Brattico, E., & Jacobsen, T. (2010). Aesthetic judgments of music in experts and laypersons—An ERP study. *International Journal of Psychophysiology, 76,* 40–51. doi:10.1016/j.ijpsycho.2010.02.002

Myford, C. M. (1989). *The nature of expertise in aesthetic judgment: Beyond inter-judge agreement*. Unpublished doctoral dissertation, University of Georgia, Athens.

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. doi:10.1037/a0012839

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology & Marketing, 26,* 470–478. doi:10.1002/mar.20283

Rawlings, D., Barrantes-Vidal, N., & Furnham, A. (2000). Personality and aesthetic preference in Spain and England: Two studies relating sensation seeking and openness to experience to liking for paintings and music. *European Journal of Personality, 14,* 553–576. doi:10.1002/1099-0984(200011/12)14:6<553::AID-PER384>3.0.CO;2-H

Reiter-Palmon, R., Mumford, M. D., Boes, O. J., & Runco, M. A. (1997). Problem construction and creativity: The role of ability, cue consistency, and active processing. *Creativity Research Journal, 10,* 9–24. doi:10.1207/s15326934crj1001_2

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach,* (pp. 97–133). Cambridge, MA: MIT Press.

Silvia, P. J. (2006). Artistic training and interest in visual art: Applying the appraisal model of aesthetic emotions. *Empirical Studies of the Arts, 24,* 139–161. doi:10.2190/DX8K-6WEA-6WPA-FM84

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. doi:10.1037/1931-3896.2.3.139

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review, 104,* 66–89. doi:10.1037/0033-295X.104.1.66

Simonton, D. K. (2004). Psychology's status as a scientific discipline: Its empirical placement within an implicit hierarchy of the sciences. *Review of General Psychology, 8,* 59–67. doi:10.1037/1089-2680.8.1.59

Simonton, D. K. (2009). Varieties of (scientific) creativity: A hierarchical model of disposition, development, and achievement. *Perspectives on Psychological Science, 4,* 441–452. doi:10.1111/j.1745-6924.2009.01152.x

Simonton, D. K. (in press). What is a creative idea? Little-c versus Big-C creativity. In J. Chan, & K. Thomas (Eds.), *Handbook of research on creativity*. Cheltenham Glos, UK: Edward Elgar.

Stein, M. I. (1974). *Stimulating creativity: Vol. 1. Individual procedures*. New York, NY: Academic Press.

Turner, S. A., & Silvia, P. J. (2006). Must interesting things be pleasant? A test of competing appraisal structures. *Emotion, 6,* 670–674. doi:10.1037/1528-3542.6.4.670

Voss, J. F., Wolfe, C. R., Lawrence, J. A., & Engle, R. A. (1991). From representation to decision: An analysis of problem solving in international relations. In R. J. Sternberg, & P. A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp. 119–158). Hillsdale, NJ: Erlbaum.

Received August 9, 2012
Revision received January 9, 2013
Accepted January 15, 2013 ■