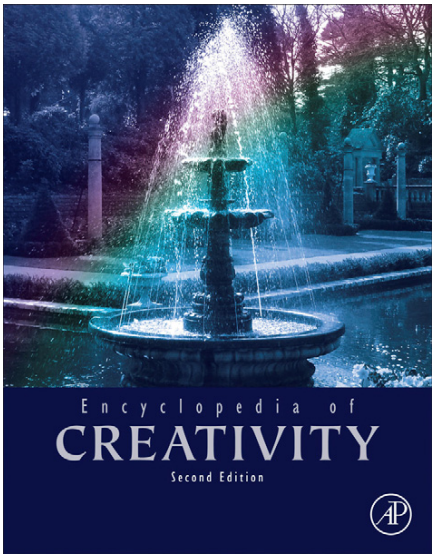

Hennessey B.A., Amabile T.M., and Mueller J.S. (2011) Consensual Assessment. In: Runco MA, and Pritzker SR (eds.) Encyclopedia of Creativity, Second Edition, vol. 1, pp. 253-260 San Diego: Academic Press.

Consensual Assessment

B A Hennessey, Wellesley College, Wellesley, MA, USA
T M Amabile, Harvard Business School, Boston, MA, USA
J S Mueller, University of Pennsylvania, Philadelphia, PA, USA

This article is a revision of the previous edition article by Beth A Hennessey and Teresa M Amabile, volume 1, pp. 347–361, © 1999, Elsevier Inc.

Glossary

Conceptual definition of creativity A product is considered creative to the extent that it is both a novel and appropriate, useful, correct, or valuable response to an open-ended task.

Construct validity The strength of the link between the term used to refer to a particular phenomenon or construct (e.g., 'creativity') and the actual features of the behavior or outcome being measured (e.g., 'degree of novelty,' 'degree of appropriateness'). Considerations of construct validity are sometimes further broken down into questions concerning both the predictive and the concurrent validity of a measure.

Convergent validity A means of establishing a test's validity by demonstrating the degree of relationship between a variety of measures of the same construct.

Ecological validity The generalizability of an experimental result to a relevant real-world population, setting, or situation.

Operational definition of creativity A product or response is considered creative to the extent that appropriate observers independently agree that it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated.

Reliability The reliability of a measure involves its consistency. In the case of the consensual assessment technique, reliability is measured in terms of the degree of agreement among raters as to which products are more creative, or more technically well done, or more aesthetically pleasing than others.

Validity The validity of a test or procedure refers to whether it is measuring what it is purported to measure.

Consensual Assessment is a technique used for the assessment of creativity and other aspects of products, relying on the independent subjective judgments of individuals familiar with the domain in which the products were made.

Introduction

Creativity is a concept that is difficult to define and even more difficult to measure. Since the Encyclopedia of Creativity was first published in 1999, the field of creativity research has seen a gradual shift away from an almost exclusive emphasis on the creative person towards a more balanced inquiry that centers both on individual difference issues and questions about the nature of creative products and the conditions that facilitate the creation of those products. But how are we to decide whether one product is more creative than another? Is it appropriate for such creativity criteria to be laid out by the researcher? Or perhaps the creators themselves should have the final say? The consensual assessment technique (CAT) for assessing creativity is based on the assumption that a panel of independent raters familiar with the product domain, persons who have not had the opportunity to confer with one another and who have not been trained by the researcher, are best able to make such judgments. Over 30 years of research have clearly established that product creativity can be reliably and validly assessed based upon on the consensus of experts. Although creativity in a product may be difficult to characterize in terms of specific features, it is something that people can recognize and agree upon when they see it.

The CAT has been successfully used in hundreds of between-subjects designs focused on the question of whether some conditions are more conducive (or detrimental) to creativity than others. The conditions to be experimentally compared can be naturally occurring (as in field studies conducted in industrial/organizational studies) or artificially created and manipulated by an experimenter in a laboratory setting. Study participants can be drawn from a single, underlying population or they might represent persons coming from different backgrounds, cultures, etc. Within-subjects designs can also incorporate the CAT to explore whether some conditions are especially conducive to or detrimental to creativity. And, again, these conditions could be naturally occurring or manipulated in the laboratory. In addition, the CAT can be used to compare product ratings made by different groups of judges. Ratings made by experts in a field can be compared to ratings made by novices. Ratings made by supervisors in the workplace or teachers in schools can be compared to ratings made by employees or students. Ratings made by children can be compared to ratings made by adults. And ratings made by judges in one culture could be compared to ratings made by judges in another culture.

The Unique Assessment Concerns of Creativity Researchers

Many empirical investigations of personality traits or cognitive styles associated with creative performance employ some form of paper-and-pencil creativity test. A variety of personality checklists, developed by Gough, Torrance, Cattell, and

others, have often been used to identify highly creative persons; however, some creativity indices have focused on behavioral factors. These behavioral assessments, such as the Torrance Test of Creative Thinking (TTCT, also known as the Minnesota Test of Creative Thinking), have typically built on Guilford's theory of divergent thinking; they elicit oral, written, and drawn (nonverbal) responses from participants.

What does it mean when someone scores high (or low) on these creativity tests? Should high scorers be considered 'creative persons?' Many creativity measures might accurately tap one or more creative abilities or predispositions, but it is most unlikely that a single test could be developed that would capture the full range of creativity components. Also troublesome is the fact that a variety of social and environmental factors have been found to influence test results. A number of studies have revealed that study participants' scores can be improved simply by telling them that creative responses will be valued. Testing environments can also influence test outcomes, and many investigations have shown variability in creativity test scores under different testing conditions and time constraints.

Even if these contextual and situational factors could be controlled for, the construct validity of many of these tests has been seriously questioned, as has the convergent validity of different test procedures considered together. This validity issue is especially problematic given the fact that many of the leading creativity tests have been validated against one another. Finally, one additional concern involves the fact that while the scoring procedures utilized in many of the creativity tests are purported to be objective, performance is often rated according to criteria based upon the test constructor's own, intuitive notion of what is creative.

Early Applications of Consensual Assessment

Mindful of these and other difficulties inherent in the creativity testing process, a number of researchers have chosen to follow a very different path. It is this group's conviction that creativity judgments can ultimately only be subjective. Rather than attempting to objectify the creativity rating process, these investigators rely on the consensual assessment of persons or products. Although, in the past, this approach was used much less frequently than creativity tests, the subjective assessment process has a long history. As early as 1870, Galton was relying on biographical dictionaries to select outstanding literary men and scientists – a technique that depended on both the subjective assessment of Galton and those who had compiled the dictionaries. Castle also used biographical dictionaries to construct an initial sample of subjects for a study of highly accomplished men and women and Cox drew her pool of geniuses for a personality study from a list of the 1000 most eminent individuals in history that had been formulated by Cattell. More recently, Simonton, in studies of sociocultural influences on creativity, developed a measure of creativity based on frequency of citation in histories, anthologies, and biographical dictionaries.

Other investigations have relied on the judgments of a select group of experts to assess the creativity of particular individuals. For example, an expert-nomination procedure was carried out by MacKinnon and his colleagues for a series of studies in the 1960s at the Institute for Personality Assessment and Research in Berkeley, California. In order to gather their subjects, these researchers asked the dean and four colleagues at the College of Architecture at the University of California to list and rate the 40 most creative architects in the United States. Similarly, Helson and Crutchfield gathered mathematicians' nominations for the most highly creative women in their field; and Barron requested that three professors of English and one editor of a literary review suggest names of creative writers.

Shifting their focus away from the creativity of persons, some researchers have asked raters to make assessments of the creativity of particular *products*. In the majority of investigations of this type, the researcher has either presented judges with his own definition of creativity for them to apply or has trained them beforehand to agree with one another. While such methodologies may successfully avoid many of the problems inherent in paper-and-pencil creativity tests, the fact that judges have been carefully instructed in the rating process calls into question both the claim of judge-based subjectivity and the meaning of interjudge reliability. Rather than impose specific definitions of creativity or related dimensions, researchers would be better served if they allowed judges to make their own, independent product assessments. In this way, creativity assessments will more closely mirror real-world assessments.

In 1976, Getzels and Csikszentmihalyi did just this when they requested that four different groups of judges (two expert and two nonexpert) use their own individual criteria when rating subjects' drawings on originality, craftsmanship, and overall aesthetic value. Sobel and Rothenberg (1980) also utilized this subjective assessment technique when they asked their raters, two accomplished artists, to judge sketches on originality, value, and overall creative potential guided only by their own subjective definitions of these dimensions.

Investigations such as the ones described above managed to overcome much of the criticism levied against the earliest applications of consensual assessment to product creativity, yet a variety of difficulties still remained. First, many of the procedures being utilized failed to differentiate between the creativity of products and other related constructs such as technical correctness or aesthetic appeal. Further, most researchers using consensual assessment procedures did not clearly state an operational definition of creativity in their publications, even when they had trained their judges to recognize specific creativity criteria in products. Nearly all contemporary definitions of creativity are conceptual rather than operational. They were never intended to be translated into actual assessment criteria. Either investigators failed to explicitly state the definition of creativity guiding their research or they presented conceptual definitions that did not adequately reflect the rating procedures they had chosen to utilize.

Systematizing the Consensual Assessment Technique in Creativity Research

The consensual assessment of creativity was formalized and systematized by Amabile's work in the social psychology of creativity, beginning in the late 1970s. When this program of

investigation was begun over 30 years ago, existing creativity measurement tools, including available subjective assessment methodologies, could not meet the unique research requirements of investigators interested in the social psychology of creativity. The majority of available assessment techniques resembled personality or IQ tests, in that they viewed creativity as an enduring personality trait. Whether they requested that a picture be completed, unusual uses for a brick be generated, adjectives describing the self be selected, or remote associations be discovered, most paper-and-pencil measures had been specifically constructed to maximize individual differences. Even existing subjective assessment methods relied on products or entire bodies of work that depended heavily on an individual's level of expertise. Prior methods had been constructed to do exactly what social psychologists try to avoid.

Social psychologists often investigate the effect of the social environment on a person's motivation for creativity, assessing both motivation and product creativity. Effects of the social environment on product outcomes are best revealed when individual difference effects are minimized. In other words, creative performance on the task must not depend heavily on participants' specialized skills. For this reason, social psychological research requires either that the task not depend heavily on special skills, or that all study participants have roughly the same skill level. If these requirements cannot be met, then the researcher should at least assess initial skills levels so they may be controlled for in analyses.

Prior to Amabile's work, the literature had not identified a methodology that could de-emphasize individual differences between subjects. In addition, researchers had not agreed upon and consistently employed an operational definition of creativity. Amabile's first step was to adopt two complementary definitions of creativity: an underlying conceptual definition to use in building a theoretical formulation of the creative process and an operational definition to apply in empirical research.

Amabile developed the following conceptual definition of creativity: a product or idea is creative to the extent that it is a novel and appropriate response to a heuristic task. This definition is similar to a number of others that came before it. Despite the implicit emphasis on the person in creativity assessment, most explicit definitions have used the creative product as the distinguishing sign of creativity. Indeed, the criteria of product novelty and appropriateness have long been seen as the hallmarks of creativity by a number of theorists.

The CAT is grounded in the original operational definition that Amabile developed: a product or response is creative to the extent that appropriate observers agree it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated. Importantly, this consensual definition is based on the creative product rather than the creative process. In fact, the majority of creativity assessment techniques require that subjects produce something – a list of ideas, a series of pictures, or the like. What does set this methodology apart from the rest is that, rather than responding to a series of predetermined items or questions, subjects simply produce an actual product such as a poem, a collage, or a story.

Perhaps the most important feature of this consensual definition is its reliance on subjective criteria. In this way, it overcomes the difficulty of attempting to specify 'ultimate' objective criteria for identifying products as creative. Indeed it may be impossible to articulate such ultimate criteria. Just as the judgment of attitude statements as more or less favorable or the identification of individuals as 'physically attractive' depends on social context, so too does the judgment of creativity. Certainly, there must be particular characteristics of attitude statements or persons or products that observers systematically look to in rating them on scales of favorability or physical attractiveness or creativity, but in the end the choice of these characteristics is a subjective one. Important writings by Gardner, Csiksentmihalyi, and others argue that creativity arises from a combination of three sources: a cultural/historical context that imposes specific symbolic rules on the creator, the creator who introduces novelty into that symbolic domain, and a field of experts who point out and validate the creative accomplishment. Thus, all judgments of creativity are necessarily relative and bounded by time and place. Creativity should not be seen as residing inside the head of the artist or scientist. Nor does it reside in a particular culture or time period or with judges representing a field of expertise within a particular era or culture. Instead, creativity must be seen as the result of a complex interaction between these three components of creator, domain, and field.

Amabile and her colleagues have attempted to capture the essential characteristics of the conceptual and operational definitions of creativity in the CAT as used in experimental studies of creativity. First, subjects are presented with tasks that leave room for considerable flexibility and novelty of response (open-ended, heuristic tasks). Second, these are tasks for which the range of appropriate responses has been clearly identified in subjects' instructions. Finally, in employing the CAT, researchers do not impose on raters their own specific views of what is creative or allow raters to influence each other; rather, raters work independently and are guided by their individual subjective conceptions about creativity.

Refining the CAT

The CAT rests on two important assumptions. First is the assumption that it is possible to obtain reliable judgments of product creativity, given an appropriate group of judges. In other words, although creativity in a product may be difficult to characterize in terms of specific features, it is something that people can recognize when they see it. Furthermore, people familiar with such products can agree with one another on this perception. A second assumption is that there are degrees of creativity such that observers can say, at an acceptable level of agreement, that some products are more or less creative than others.

Procedural Requirements

Researchers deciding to utilize the CAT should make certain that a number of requirements are met. First, the judges should all have had some experience (and roughly equivalent experience) with the domain in question. When Amabile was first developing the CAT, she and her colleagues sometimes relied

on the notion of 'expert' to describe an appropriate body of raters. Over the years, extensive work with this methodology has, however, brought about a tempering of this view. Basically, the method requires that all those rating products be familiar enough with the domain to have developed, over a period of time, some implicit criteria for creativity, technical goodness, and so on. For example, when asked to rate the creativity of paper collages, both children and adults from a variety of backgrounds have produced highly reliable assessments. When dealing with a more specialized and esoteric field, such as physics or computer programming, however, the range of 'experts' (i.e., appropriate observers) would certainly have to be considerably narrower. In either case, it is the judges' familiarity with the domain that is important, not the fact that they, themselves, may have produced work rated as highly creative.

A second requirement is that the judges must make their assessments independently. They are not trained by the experimenter to agree with one another; are given no specific criteria for judging creativity; and are not allowed to confer in their assessments.

Third, judges should be instructed to rate the products relative to one another, rather than rating them against some absolute standards they might hold for drawing, sculpture, poetry, and so on. This is important because, for most studies, the levels of creativity produced by the 'ordinary' subjects who participate will be very low in comparison with the greatest works ever produced in that domain.

Fourth, each judge should view the products in a different random order. If all judgments are made in the same order by all raters, high levels of agreement might reflect methodological artifacts.

Finally, if this technique is to be used to evaluate performance on a task to which it has not been applied in the past, judges should be asked to rate the products on other dimensions in addition to creativity. Minimally, they should make ratings of technical aspects of the work, and if appropriate, its aesthetic appeal as well. These additional assessments make it possible to examine the degree of relatedness or independence of these dimensions in subjective judgments of the products in question.

Once the judgments are obtained, ratings on each dimension should be analyzed for interjudge reliability. In addition, if several subjective dimensions of judgment have been obtained, these should be entered into a factor analysis to determine the degree of independence (discriminant validity) between creativity and the other dimensions investigated. Finally, if the products lend themselves to a straightforward identification of specific objective features, these features may be assessed and correlated with creativity judgments. Prior studies have identified some physical features of collages and some verbal features of stories that correlate with creativity judgments.

Reliability

Given the consensual definition of creativity, the most important criterion for the results of this assessment procedure is that the product ratings be reliable. In order to compute reliability, Amabile originally utilized the Spearman–Brown prediction formula that is based on the number of judges (n) and the mean interjudge correlation (r):

$$\text{reliability} = \frac{nr}{1 + (n-1)r}$$

This technique yields results highly similar to the Cronbach's coefficient alpha as calculated by the 'reliability analysis' procedure in the Statistical Package for the Social Sciences (SPSS). In the interest of simplicity, in recent years researchers employing the CAT have relied upon the SPSS calculation as their measure of interrater agreement. In most instances, a reliability figure of 0.70 or higher can be considered evidence of an acceptable level of agreement between judges. Once such a level is reached, it is then appropriate to compute a sum (or an average) across all ratings given each product. These sums (or averages) then constitute the unit of analysis for further computations.

By definition, interjudge reliability in this method is equivalent to construct validity: if appropriate judges independently agree that a given product is highly creative, then it can and must be accepted as such. In addition, it should be possible to separate subjective judgments of product creativity from judgments of technical goodness and aesthetic appeal. Within some domains, it may be difficult to obtain ratings of product creativity that are not highly positively correlated with judges' assessments of product technical goodness or aesthetic appeal. Yet it is essential to demonstrate that it is at least possible to separate these dimensions, otherwise the discriminant validity of the measure would be in doubt. In other words, judges might be rating a product as 'creative' merely because they like it or believe that it is technically well-done.

Supporting Data

In the program of research carried out by Amabile and her colleagues over the last 30 years, numerous studies have demonstrated that the subjective assessment technique described above does, in fact, yield reliable measurements appropriate for social psychological studies of creativity. In studies employing a paper collage task, participants are presented with a piece of cardboard, glue, and a variety of colored pieces of paper of different shapes and sizes. They are instructed to make a design that 'makes them feel silly,' and they are given approximately 15 minutes to engage in the task. In the majority of instances, professional artists and/or graduate students in the studio arts have served as the 'expert' judges. In those investigations enlisting elementary or preschool students as participants, classroom art teachers familiar with the work of children have also been recruited. For collage ratings, ten or so judges have typically been employed. Without exception, raters have yielded highly reliable assessments of collage creativity.

Equally important as interrater reliability is the requirement that judges' assessments of certain additional product dimensions do not correlate highly with their ratings of creativity. Here too the results have been very encouraging. In keeping with most theorists' conceptions of creativity, ratings of novelty and originality have typically been highly related to ratings of creativity, while ratings of various aspects of collage technical goodness have not usually been significantly correlated with creativity assessments.

In addition to the collage measure, Amabile and her colleagues have also employed a wide variety of other creativity tasks in their investigations. In an attempt to assess the impact of social constraints on verbal creativity, they have, for example, asked adults to complete five-line American haiku poems. In an effort to reduce product variability and make the judging task somewhat more manageable, study participants are typically provided with the first line of the poem they are to write. In one study, this technique was successfully adapted for use with young children. Sitting in front of a computer screen, elementary school students were prompted in a question and answer interactive format to enter one-, two-, or three-word lines. Other measures of verbal creativity that have also proven useful involve completing sentences; writing essays, descriptive paragraphs, and free-form poems; coming up with captions for cartoons; and telling a story to accompany an open-ended picture book without words. This story-telling task has been used successfully with children as young as first grade. Study participants look through the book with the experimenter and then are asked to tell a story by saying 'one thing' about each page.

Each of these verbal tasks has also yielded highly reliable creativity assessments. Whether they are poets rating haikus, elementary school teachers rating children's stories or graduate students rating cartoon captions, judges show consistently high interrater agreement.

In addition to measuring artistic and verbal performance, Amabile and colleagues have also used some creative problem-solving tasks. One assessment procedure taps spatial-mathematical creativity in children and calls for the construction of a geometric design on a computer screen. Another activity requires that young subjects fill in the outline of a geometric shape with colored pieces of felt. Problem-solving tasks involving adult subjects include the construction of computer programs, building structures from ordinary materials, generating survival ideas or ideas for high-tech products and coming up with business solutions. Although none of these techniques has been tested to the same extent as the collage-making or many of the verbal creativity tasks, it is encouraging that judges have rated products produced by children and adults with high levels of reliability.

Recent work by Hennessey and colleagues investigated whether the CAT would also produce valid and reliable creativity ratings in non-Western cultural settings. Specifically, one study recruited school teachers from the United States, Saudi Arabia, China, and South Korea to assess collages and stories created by children living in their local area. Results confirmed that across all four cultural contexts, judges' ratings of product creativity showed high levels of interjudge agreement. This suggests that the CAT is especially useful for cross-cultural investigations. Rather than impose a paper-and-pencil measure and scoring criteria originally developed for use in the West, the CAT allows for the subjective assessment of products by judges who come from the same cultural background as the study participants who produced the products.

Clearly, the CAT has wide-range application. It has been successfully employed with both child and adult subjects and allows for the assessment of creativity in a number of different domains. Over the years, subject populations have been expanded beyond the original pool of undergraduates and elementary school children, demonstrating that the creativity of professional artists, professional art students, computer programming students, student poets, and employees of a high-tech company can also be reliably assessed. Most recently, the CAT has also been demonstrated to yield reliable assessments of product creativity across a wide range of cultural contexts. For these reasons, an ever-growing number of researchers have come to rely on this assessment technique.

Taking a Closer Look

The CAT was originally developed to yield reliable measures of the creativity of products produced in experimental studies of social-psychological influences on creativity. More recently, this methodology has also been applied to field studies in organizational settings. Whether they are asked to make a collage, tell a story, write a haiku poem, or come up with new ideas for products or solutions to problems, participants in these investigations are primarily engaged in behaviors resulting in what has been termed 'little c,' 'everyday,' creativity. The CAT has consistently been shown to yield reliable measures of product creativity in these contexts, but what is it exactly that judges are doing when they set out to make these ratings? Are they assessing only the completed product or are they also making assumptions about the process that went into producing that product? Can judges be expected to reliably assess features of the creative process? In 1994, Hennessey conducted a series of four studies with these questions in mind.

In the first of these investigations, undergraduate students rated either the creativity, technical goodness and likeableness of geometric line designs that had been created on a computer or they rated computerized replays of the procedure that went into producing each of these products. Reliability was high and raters who had been asked to make assessments of process had no more difficulty than did raters assessing finished products. A strong and positive relation was found between ratings of product creativity and ratings of the creativity of the processes that went into completing those finished products. And similar strong and positive correlations were found between ratings of the technical goodness of finished products and ratings of the technical goodness of the processes that lead to the completion of those products. In Study II, a separate group of undergraduate students made assessments of both process and finished products. Reliability was again acceptable. Judges found the rating of process no more difficult than the rating of finished products, and their ratings of process and product were positively correlated. A third study then explored whether these same results would obtain when 'real-world' drawings produced by Pablo Picasso were assessed. Undergraduates in this investigation rated videotaped segments of the processes that went into completing four Picasso drawings and stills of those drawings taken from the movie 'The Mystery of Picasso.' Importantly, these videotaped segments and stills had all been pretested on another sample of undergraduates, and not a single study participant had guessed that the drawings had been done by Picasso, or any other 'big-C' artistic master for that matter. These products were not typical of Picasso's work and were utilized simply as a matter of convenience, because a professionally-produced video of the process that

went into creating the drawings was readily available. Reliability among judges was again acceptable and correlations between ratings of process and product were of approximately the same magnitude as those obtained in the two previous investigations.

But what are judges actually doing when they make their ratings? When asked to assess product creativity, are they considering only the final product? Or do they also take into account other factors – factors such as information about the circumstances under which a product was produced or the characteristics of the creator? In investigations employing the CAT, judges are typically given very little information about the persons who have made the products they are to rate. Most often, they are instructed in the assessment process and are told simply that the materials they will be viewing were produced by university undergraduates, or preschoolers, or some other group. Implicit in this procedure is the assumption that creativity is a unitary construct independent of factors such as background or experience of the creator.

The last of Hennessey's four investigations was intended as a preliminary exploration of the impact of artist age information on judges' creativity assessments. One group of undergraduates was asked to judge collages made by children and adults after receiving accurate information about the age of the artists. A second group was asked to rate the same collages after receiving false, reversed information as to the age of the artists. Finally, a third group of undergraduates was asked to judge the collages without being given any information as to the age of the artists. Reliabilities were highly acceptable for all three of the judgment conditions, and age information was found to have a significant effect. The highest creativity ratings were given to adults' collages that had been falsely labeled as children's products. The lowest creativity ratings were given by judges who had received no age information to collages that had been produced by children. Overall, it was found that those raters receiving age information about the artists, whether accurate or reversed, gave products higher ratings of creativity than did raters for whom no age information was available. Within age information groups, no significant differences emerged between judges' creativity ratings of children's and adults' collages.

Six of the 33 judges polled reported that they had considered artists' ages when making their product assessments. Two other respondents mentioned 'fighting' against the tendency to take artist age into consideration. Contrary to expectation, it was the mere availability of age information and not the specific adult or child label that affected raters' judgments. Whether raters were given an accurate or a reversed age label, they judged children's collages to be higher in creativity than did raters given no age information. This finding suggests that creativity theorists and other researchers wishing to employ the CAT must be certain to note whether age information has been made available, either purposefully or unintentionally, to judges. Similarly, careful assessments should be conducted to determine whether raters have made any age inferences on their own.

Does knowing a subject's identity inject bias into judges' creativity ratings? This question is particularly pertinent to field studies of creativity within the realm of organizational behavior, as they frequently elicit supervisory ratings of subordinates' creative behavior. This procedure necessitates that supervisors know the identity of the individuals who have produced the products or generated the ideas they are rating. While these studies typically control for workers' gender, age, organizational tenure, and other demographic variables that could impact creativity ratings, research has yet to determine the extent to which supervisors' attitudes, affective states or biases – such as feelings of liking and favoritism – might impact their ratings in significant ways. This issue is of particular concern because most organizational studies rely on the ratings of only a single supervisor, violating a central tenet of the CAT. However, a recent study by Baer and Oldham did find that independent ratings by two supervisors correlated highly. Of course, it is possible that both supervisors were subject to the same biasing forces. Thus, ratings of an individual's work by people who know that individual must always be regarded with caution.

Another related issue is whether individuals can make reliable ratings of the creativity of their own work. Researchers have typically found moderate correlations between creativity self-assessments and mean ratings made by others, although self-ratings often show a positivity bias. For example, in a recent study by Moneta and colleagues, workers' monthly self-assessments of their creative contributions to a project correlated moderately and significantly with independent ratings made by coworkers and supervisors. However, the mean self-ratings were higher than either the mean coworker ratings or mean supervisory ratings. Other researchers have found moderate correlations between self-ratings and peer-ratings. Indeed, a recent meta-analysis conducted by Heidemeier and Moser yielded an overall correlation of 0.22 between self and supervisory performance ratings. In sum, the literature suggests that self-ratings of creativity do relate to ratings made by external observers, but may be biased toward positivity.

A variety of papers have carefully explored these and other issues concerning the question of who should be considered an appropriate judge. In one study, parents and teachers were found to be equally accurate at recognizing the creativity of children's ideas. But other investigators found that young children's judgments about art were considerably different from those offered by older children. Runco and colleagues asked college students to each produce three three-dimensional artworks that were then rated by the subjects themselves, a group of their peers, and three professional artists. Analyses revealed that the student subjects saw significant differences in the creativity of their own three art projects. Similar differences also were reflected in the peer ratings of the artwork. The assessments made by the professional artists, however, failed to reflect significant differences in creativity between products. Thus, vast differences in level of expertise between study participants and judges may influence creativity judgments.

Dollinger and Shafran reopened the question of whether nonexpert judges might reliability rate the creativity of drawings made by a sample of nonprofessional artists. As mentioned previously, the CAT requires that researchers refrain from training judges so as not to impose their own views and risk shaping judges' ratings. However, in an interesting modification of the CAT, Dollinger and Shafran calibrated nonexpert judges' artistic creativity ratings by exposing them to 16 prototype drawings and corresponding ratings made by expert

judges in a prior experiment. Subsequent to the calibration, the mean correlation between expert and nonexpert ratings for a second set of products was 0.91.

Taken together, these results suggest that some clarification or modifications be made to the CAT specification about what qualifies as an appropriate level of judge expertise. When rating products produced by either nonprofessional or professional individuals, appropriate judges should be defined as persons whose expertise matches or exceeds the expertise of those individuals who created the products. In the case of products produced by nonprofessionals, if researchers desire nonexpert judges' ratings to correspond to ratings made by expert judges, researchers may use a calibration technique. Finally, if creativity assessment procedures require raters to have familiarity with the entire body of a given person's work – as is often the case in field research – the most appropriate judges will be those with the greatest knowledge of the subject's performance and output. Importantly, as cautioned earlier, future research should consider the extent to which knowledge about a creator's identity might bias the validity of such ratings.

Some Recent Developments

Over the years, the CAT has come to serve as an invaluable tool for a number of creativity researchers. This methodology has been extended to a variety of tasks in a variety of domains, and the diversity of subject populations and rater populations being studied is also constantly growing. Researchers have shifted from employing the CAT exclusively in tightly controlled experimental settings, to also using the CAT in quasiexperimental or field studies in which participants create products under a variety of different conditions and situations. Perhaps the most prolific expansion of the CAT in the recent decade has occurred in the organizational domain. In field studies designed to investigate the creative performance of professionals, supervisors and/or coworkers are frequently called upon to serve as raters. As noted earlier, the majority of organizational creativity studies rely on the ratings of a single supervisor per study participant. This dependence on supervisor ratings draws on a decades-long tradition in the organizational literature of using such assessments to obtain quantitative measures of an employee's performance. Although the traditional CAT involves multiple judges rating the creativity of products using a single-item scale ('creativity'), the organizational domain has adapted the CAT to involve a single supervisor who rates several subjects on a multiitem scale (e.g., 'Searches out new technologies, processes, techniques, and/or product ideas'; 'Generates creative ideas').

Investigations carried out in the organizational domain pose unique challenges that make strict adherence to the CAT protocol difficult, if not impossible. Most prominently, while it would be preferable to obtain independent creativity ratings from multiple judges, it is often difficult to find more than one person who has access to the same range of information about the work done by a given employee. Moreover, because employees often work in team contexts, the identification of a single individual's contribution to a creative product developed by a team requires knowledge about the specific set of tasks to which the individual was assigned. Due to intense pressures in organizations, higher-level managers often do not have the time to interact with subordinates and instead rely on immediate supervisors to provide performance evaluations and feedback. Indeed, Hoegl and Gemuenden suggested that ratings made by higher-level managers reflect external market pressures more than they do the actual performance and functioning of team members. Thus, higher-level managers are seen as unsuitable judges of the creativity of most employees' work.

To circumvent these difficulties, some organizational research has employed peer ratings of creativity – requiring every member of a team to independently rate the performance of every other member. This method appears particularly promising, as it allows researchers to assess interjudge reliability. Moreover, research suggests that peer ratings are highly correlated with supervisory ratings of a target's performance, which, in turn, correlate significantly with some objective measures of creativity such as invention disclosures. The research evidence clearly suggests that it is preferable to use multiple raters (i.e., supervisors and coworkers, if available) to improve the reliability of creativity assessments in organizational studies. Nonetheless, perhaps for the sake of expediency, most organizational creativity researchers have tended to rely on ratings made by a single supervisor.

One additional arena of organizational creativity research that now frequently employs the CAT is the research on group creativity. Since the 1999 publication of this encyclopedia, the investigation of group creativity has blossomed into a distinct, multifaceted, and highly prolific field of inquiry. Early brainstorming studies in the laboratory typically employed the CAT, requiring three or more expert raters to judge the originality of each idea produced by a group. The developing body of group creativity field research – like field research on individual-level creativity – has measured creativity in a variety of ways. Some field studies have employed a single rater to assess a specific product. At other times, group creativity research has employed multiple raters assessing either a single product or the holistic creativity of a group. This CAT protocol allows for the calculation of reliability estimates, and is thus preferable to employing single judges.

Conclusions

Clearly, the CAT has been a great boon to many creativity researchers. It has broad application, is founded on a clear operational definition, and can be adapted to suit a wide variety of research situations. Moreover, with its similarity to real-world creativity judgments, the CAT enjoys a high degree of ecological validity. Despite these advantages, however, the CAT should not be considered an ultimate and universally useful means of creativity assessment. Indeed, this assessment methodology has some specific limitations. Most notably, if time concerns are paramount, this approach is decidedly impractical. Choosing an appropriate task as well as an appropriate body of judges can be extremely time-consuming, as can the assessment of products and the necessary statistical data analyses. However, a number of creativity researchers continue to believe that the benefits of the CAT outweigh its costs; and recently, Kaufman and colleagues successfully experimented

with a modified CAT technique that reduced time demands yet still yielded reliable assessments.

Perhaps the greatest strength of the CAT rests in the flexibility it affords to creativity researchers. First, the CAT can be used to obtain reliable assessments of the relative creativity (technical goodness, aesthetic appeal, etc.) of products made by a variety of individuals. Second, the CAT can be expanded to new subject populations, new performance domains, and new tasks that are quite different from those originally envisioned. In mimicking the way in which creativity is judged every day in the arts, the sciences, and the professions, the CAT helps bring creativity from the realm of the mysterious and the mystical, where it remained for centuries, into the realm of the understood and the accessible.

See also: Creative Products; Definitions of Creativity; Divergent Thinking; Everyday Creativity; Historical Conceptions of Creativity; Novelty; Pablo Picasso 1881–1973; Research: Quantitative; Social Psychology; Testing/Measurement/Assessment.

Further Reading

Amabile TM (1982) The social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43: 997–1013.

Amabile TM (1996) *Creativity in Context*. Boulder, CO: Westview Press.

Baer J, Kaufman JC, and Gentile CA (2004) Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal* 16: 113–117.

Baer M and Oldham GR (2006) The curvilinear relation between experienced time pressure and creativity: Moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology* 91: 963–970.

Dollinger SJ and Shafran M (2005) Note on consensual assessment technique in creativity research. *Perceptual and Motor Skills* 100: 592–598.

Hennessey BA (1994) The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal* 7: 193–208.

Hennessey BA, Kim G, Zheng G, and Sun W (2008) A multi-cultural application of the consensual assessment technique. *The International Journal of Creativity and Problem Solving* 18: 87–100.

Hoegl M and Gemuenden HG (2001) Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science* 12: 435–449.

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

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

Moneta GB, Amabile TM, Schatzel EA, and Kramer SJ (2010) Multi-rater assessment of individual creative contributions to team projects in organizations. *European Journal of Work and Organizational Psychology* 2: 150–176.

Plucker JA and Runco MA (1998) The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. *Roeper Review* 21: 36–39.

Shalley CE, Zhou J, and Oldham GR (2004) The effects of personal and contextual characteristics on creativity: Where should we go from here? *Journal of Management* 30: 933–958.

Simonton DK (2007) Historiometrics. In: Salkind NJ (ed.) *Encyclopedia of Measurement and Statistics*, vol. 2, p. 441. Thousand Oaks, CA: Sage.

Sobel RS and Rothenberg A (1980) Artistic creation as stimulated by superimposed versus separated visual images. *Journal of Personality and Social Psychology* 39: 953–961.

Zhou J and Shalley CE (eds.) (2008) *Handbook of Organizational Creativity*. New York: Lawrence Erlbaum Associates.

Relevant Websites

http://www.ccl.org/leadership/index.aspx – Center for Creative Leadership.

http://www.informaworld.com/smpp/title~content=t775653635~db=all – Creativity Research Journal.

http://www.apa.org/about/division/div10.html – Society for the Psychology of Aesthetics, Creativity and the Arts.