Creativity Research Journal
2004, Vol. 16, No. 1, 113–117

Extension of the Consensual Assessment Technique to Nonparallel Creative Products

John Baer
Rider University

James C. Kaufman
California State University at San Bernardino

Claudia A. Gentile
Educational Testing Service

ABSTRACT: The consensual technique for assessing creativity is widely used in research, but its validation has been limited to assessing the creativity of artifacts produced under tightly constrained experimental conditions. Typically, only artifacts produced in response to very similar instructions have been compared. This has allowed researchers to compare such things as the effects of different motivational conditions on creative performance, but it has not allowed many other kinds of comparisons. It has also limited the use of the technique to artifacts gathered for specific experimental purposes, as opposed to already-existing artifacts produced under less controlled conditions. For this study, samples of writings collected by the National Assessment of Educational Progress that were written in response to a very wide variety of assignments and under varying conditions were rated for creativity by 13 expert judges. Judges compared the creativity of 103 stories, 103 personal narratives, and 102 poems, all written by 8th-grade students. Very high levels of interrater reliability were obtained, demonstrating that the consensual method can be validly extended to such samples. New avenues for future research made possible by these findings are then discussed.

Amabile's (1982) pioneering work developing and validating the consensual assessment technique for evaluating the creativity of diverse creative products has made possible a broad range of experimental studies of creativity. The essential features of this procedure include giving subjects some prompt or instruction for creating some kind of product, and then having experts independently assess the creativity of those artifacts. For example, in one study "students were given a line drawing of a girl and a boy … [and] asked to write an original story in which the boy and the girl played some part" (Baer, 1994a, p. 39). The experts were then asked to rate the creativity of the stories on 1.0-to-5.0 scale, based on their own expert sense of what is more or less creative. Expert judges need not explain or defend their ratings in any way. They are simply asked to use their expert sense of what is creative in the domain in question to rate the creativity of the products in relation to one another.[1]

[1]The actual instructions given to raters in the example given were: "There is only one criterion in rating these tests: creativity. I realize that creativity doesn't exist in a vacuum, and to some extent creativity probably overlaps other criteria one might apply—aesthetic appeal, organization, richness of imagery, sophistication of expression, novelty of word choice, appropriateness of word choice, and possibly even correctness of grammar, for example—but I ask you to rate the stories solely on the basis of your thoughtful-but-subjective opinions of their creativity. The point is, you are the expert, and you needn't defend your choices or articulate a definition of creativity. What creativity means to you can remain a mystery—what I want you to do is use that mysterious expert sense to rate the stories for creativity" (Baer, 1994a, p. 39–40).

The authors thank Kathy Howell, Venus Mifsud, Susan Martin, and Alyson Tregidgo for their assistance, and Fred Cline for help with data analysis. This article was supported by a grant from the National Center for Educational Statistics.

Correspondence and requests for reprints should be sent to John Baer, Memorial Hall, Rider University, Lawrenceville, NJ 08648. E-mail: baer@rider.edu.

The consensual assessment technique is both widely used and well validated in creativity research. It has been employed in diverse experiments using a wide range of tasks (e.g., writing poems and stories, telling stories to go with pictures, creating collages and other artworks, and creating mathematical word problems and puzzles) with both children and adults as subjects. In study after study, these expert ratings, done completely independently of one another and without rubrics of any kind, have yielded quite satisfactory inter-rater reliabilities, with coefficient alphas that typically exceed .70 and often range as high as .90 or higher (Amabile, 1983, 1996; Baer, 1993, 1998a; Hennessey & Amabile, 1999; Runco, 1989).

In these research studies, the instructions given to subjects are typically quite uniform. In a collage-making task, for example, all subjects may be given identical sets of materials and instructed simply to make the "most interesting" collages they can (Baer, 1998b, p. 23). In a story-writing task, as noted above, some uniform prompt is commonly provided. And in a poetry-writing task, the topic or title of the poem may be provided (e.g., Baer, 1997), although at times subjects have been free to choose their own topic (e.g., Baer, 1994a).

Although well validated for artifacts that have been created under tightly controlled conditions, the consensual assessment method has not been well-studied when it has been applied to creative products that have been produced in response to widely varying instructions, such as stories or poems created by students in different classrooms in response to different writing assignments. The question of whether consensual assessment is a valid technique for comparing such creative products is an important one; if this technique is shown to be valid, then new areas for research could be explored, such as the comparison of the creativity of students' writing in response to different kinds of assignments. It would also allow the use of the consensual assessment technique in quasi-experimental studies using already existing data gathered under diverse conditions. This would greatly increase the possibilities for creativity research and allow researchers to make use of already existing creative products (such as stories and poems), and in that way eliminate the need to collect new samples of subjects' creative products. In many cases, this would result in a considerable savings in data-collection time and effort.

Such an extension makes sense in terms of the fundamental idea underlying this technique for assessing creativity. Consensual assessments of creative products are not linked exclusively to any particular theory of creative thinking, nor do they rely on any such theory for their validity. The only "theory" upon which they are based is the belief that experts in a given domain can recognize creativity when they see it (and if experts in some domain can't do this, then no assessment of creativity in that domain can have any meaning; Baer, 1994b). Consensual assessment by recognized experts is, of course, how creativity is typically assessed in almost all fields, even the "hard" sciences (Kuhn, 1970; Simonton, 1999), although such assessments are generally much more rich and multidimensional than is possible using simple and linear 1.0-to-5.0 scales. In such real-world assessments of creativity, moreover, there is no limitation that prevents one from comparing works that were created under diverse conditions. In fact, works produced in response to similar, and tightly controlled, experimental constraints are rarely if ever the subject of real-world expert assessments of creativity. In such cases, of course, creativity at the highest levels is being assessed, which is not the kind of creativity typically evaluated in experimental studies of creativity. The goal of this study was to gather data to see if an extension of the consensual method of assessing creativity could be reliably employed when comparing artifacts that represent more "garden-variety" levels of creativity that have been produced under nonparallel (and nonexperimental) conditions.

To assess the reliability of the consensual assessment technique across writing samples collected under diverse conditions and in response to widely varying writing assignments, we employed poems, stories, and personal narratives collected as part of the National Assessment of Educational Progress (NAEP). All were written by 8th-grade students, and no more than four came from any one classroom. Thirteen experts, working independently, evaluated the creativity of these three sets of papers. Our goal was to determine if the good inter-rater reliability that has been found when such creative products are produced under carefully controlled conditions would also be found when the creative products were much more varied in their origins.

Method

Selection of Sample

The papers were all drawn from the 1998 NAEP Classroom-based Writing Study. In that study, students in a nationally representative sample of 8th-grade students were asked to assemble folders containing two samples of their best writing. Seventeen percent of the students included poetry in their folders (416 total poems), 34% included fictional stories (840 total stories), and 48% included personal narratives (1,195 total personal narratives). Approximately 125 classrooms, representing a wide variety of demographics, participated in that study. In some classrooms, many students contributed samples based on the same assignment, but in others as many as 30 different assignments elicited the papers that students in one class selected for their folders.

For our study, we selected a subsample of each set of papers in a way that assured that all school types (rural, suburban, urban), all community economic levels, and all regions of the country were represented. We also decided that no more than one paper per student should be included in our samples, even though they might come from different categories (poetry, stories, personal narratives). A total of 103 stories, 103 personal narratives, and 102 poems were selected in adherence to these guidelines.

Creativity Assessments by Experts

There were 13 expert judges, all of whom had some experience with the writing of middle school students. Among the judges were several creative writers and editors of literary journals, creative writing teachers, and psychologists who study creativity, with roughly equal representation in each of these three areas (5 writers/editors, 4 teachers, and 4 psychologists). The backgrounds of several judges included more than one of these categories, but all of the judges fell into at least one of them, in addition to having previous experience reading and evaluating the creative writing of 8th-grade students. All 13 judges read and assessed the creativity of all of the stories, personal narratives, and poems using a 6-point scale.

Judges rated the poems, stories, and personal narratives independently. To help them with the task, judges were encouraged first to divide the papers in each group into three piles (low, medium, and high creativity) and then to subdivide each pile to create six levels of creativity. In their final ratings they were free to move papers into whichever of the six levels they deemed most appropriate, regardless of their initial rankings, and they were asked to report only their final ratings. These ratings were conducted and collected entirely through the mail. Raters did not meet or talk about their ratings with one another or with the experimenters until after all the judges' ratings had been submitted.

Results and Discussion

The coefficient alpha inter-rater reliabilities were 0.940 for the stories, 0.957 for the personal narratives, and 0.868 for the poems. This is quite high—higher, in fact, than the levels found in almost all the research studies that have employed the consensual assessment technique. This unusually high level of inter-rater reliability is probably due to the unusually wide range of creativity represented in this sample. NAEP samples come from all educational levels, so these sets of papers represented a much broader and more diverse sample than that found in most of the research studies that have used the consensual assessment technique in the past.

In the field of educational research, methods used to calculate the degree of agreement among raters frequently rely on a relatively small number of raters to score students' work, because of the high cost associated with using greater numbers of expert judges. In this study, 13 experts in creativity and creative writing assigned scores to each of the 8th graders' poems, fictional stories, and personal narratives. This is similar to the numbers of judges Amabile (1983, 1996) used in most of her validation studies, but more than the number used in many research studies. The large number of raters in this study may have resulted in a higher degree of agreement among raters than would be achieved had a smaller number of raters be used (just as a test with only 5 items will generally have a lower reliability than a similar test employing 15 items). Given the extremely high levels of inter-rater reliability achieved,

however—higher, for comparison purposes, than those found in Amabile's (1983, 1996) original validation studies using similar numbers of experts judges—it seems likely that smaller numbers of judges will work satisfactorily. This has been the case with the consensual assessment technique in general (Baer, 1993).

To further explore the degree of rater agreement, a procedure was developed that would result in a more conservative estimate of inter-rater reliability. For each of the three types of writing that were rated in the study, three pairs of raters were randomly selected. Thus, for the poetry ratings, there were three pairs of raters; for the fictional story ratings, another set of three randomly assigned pairs was created; and for personal narratives a third set of three randomly assigned pairs was created.

For poetry, the inter-rater correlation coefficients ranged from .61 to .72, for an average correlation of .66. The inter-rater correlation coefficients for the fictional story ratings ranged from .73 to .81, with an average correlation of .76. For personal narratives, the inter-rater correlation coefficients ranged from .78 to .80, with an average correlation of .79.

In the field of writing assessment, there is no official standard for acceptable rates of inter-rater agreement. Current practices in large-scale writing assessments, which strive for higher rates of inter-rater agreement, often find acceptable inter-rater correlation coefficients in the .70 to .80 range (Powers, 2000). Other studies have described inter-rater correlation coefficients in the area of .75 as "excellent," and agreements below .40 as indicating a "poor" degree of agreement (Fleiss, 1981). Landis & Koch (1977) provided the following guidelines for interpreting the strength of inter-rater correlation coefficients. A correlation coefficient of .00 to .20 represents slight agreement, a coefficient of .21 to .40 represents fair agreement, a coefficient of .41 to .60 represents moderate agreement, a coefficient of .61 to .80 represents substantial agreement, and a coefficient higher than .81 is considered almost perfect.

In the field of creativity research, correlation coefficients between .70 and .80 are believed to indicate a strong degree of agreement among raters (Amabile, 1996). In considering these various approaches to interpreting inter-rater correlation coefficients, the correlation coefficients for fictional stories and personal narratives were well within acceptable ranges. The correlation coefficients for poetry were just within acceptable ranges. Thus, even when a conservative approach to determining the degree of inter-rater agreement is used, the results of this study indicate that raters tended to agree when asked to classify pieces of writing along a continuum of creativity.

The high inter-rater reliabilities obtained demonstrate that creativity ratings based on consensual assessments by experts of artifacts gathered even under very open and uncontrolled conditions are indeed valid assessments. As mentioned previously, this opens up new avenues for creativity research.

It is important to note that what is being assessed here is the creativity of the artifacts produced, not of the creators who made them, and that an individual who received a different assignment, or had been asked to create something under different conditions, might have produced something more (or less) creative (although these assessments have been shown to have fairly high stability over time when subjects have been asked to create two stories, or two poems, in response to similar prompts but at an interval of 11 months; Baer, 1994b). The consensual assessment technique was developed by Amabile (1982) to compare parallel creative works created under different motivational constraints. This evidence that it can validly assess nonparallel creative works—that is, ones not created in response to the same assignments or prompts—means that it can now be confidently used to compare such assignments or prompts to determine, for example, which tend to produce higher levels of creative products. One very concrete example of how this might be used experimentally would be to compare the creativity of stories written in response to different instructions by the same students. This would help educators understand what kinds of assignments elicit higher levels of creativity from students.

It is also important to note that the artifacts assessed in this study were not created for this study but were already in existence. There is a wealth of potential data in the various kinds of student works collected by NAEP, as well as work produced and collected for other purposes. Our results give a green light to researchers to make full use of this potential bounty of already-collected creative artifacts to evaluate diverse hypotheses regarding the factors that may be associated with, or tend to lead to, greater levels of creative performance.

References

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

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

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

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

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

Baer, J. (1994b). Performance assessments of creativity: Do they have long-term stability? *Roeper Review, 7*(1), 7–11.

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

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

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

Hennessey, B. A., & Amabile, T. M. (1999). Consensual assessment. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity, Vol. 1* (pp. 346–359). San Diego, CA: Academic Press.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd Ed). New York: Wiley.

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

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Powers, D. E. (2000). *Computing reader agreement for the GRE Writing Assessment*. Princeton, NJ: Educational Testing Service.

Runco, M. A. (1989). The creativity of children's art. *Child Study Journal, 19*, 177–190.

Simonton, D. K. (1999). *Origins of genius*. New York: Oxford University Press.