Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability

James C. Kaufman a,*, Joohyun Lee b, John Baer c, Soonmook Lee b

a Learning Research Institute, California State University at San Bernardino, United States
b SungKyunKwan University, Republic of Korea
c Rider University, United States

Received 1 November 2006; received in revised form 22 March 2007; accepted 8 April 2007
Available online 3 May 2007

Abstract

The consensual assessment technique (CAT) is a measurement tool for creativity research in which appropriate experts evaluate creative products [Amabile, T. M. (1996). Creativity in context: Update to the social psychology of creativity. Boulder, CO: Westview]. However, the CAT is hampered by the time-consuming nature of the products (asking participants to write stories or draw pictures) and the ratings (getting appropriate experts). This study examined the reliability of ratings of sentence captions. Specifically, four raters evaluated 12 captions written by 81 undergraduates. The purpose of the study was to see whether the CAT could provide reliable ratings of captions across raters and across multiple captions and, if so, how many such captions would be required to generate reliable scores, and how many judges would be needed? Using generalizability theory, we found that captions appear to be a useful way of measuring creativity with a reasonable level of reliability in the frame of CAT.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Creativity; Assessment; Consensual assessment technique

Creativity has been measured in many ways. All methods have limitations, of course, but that is not the same as saying all creativity tests are equal. Some are more valid than others for particular purposes, and some are more valid for a wider variety of purposes. Unfortunately, some of the most valid and reliable measures of creativity are also the most difficult and expensive to use. This paper offers new evidence of validity for one type of creativity, the consensual assessment technique.

What would an ideal test of creativity look like? It would need to be reliable, of course, and yet it should be sensitive to changes (such as varying situational constraints) that might result in actual within-subject differences. It would need to be valid for the purpose of the test (and a test that is valid for multiple purposes would be a more generally useful test). The test should also not be dependent on the validity of a particular theory of creativity, especially if the test is to be used in any way for theory testing or in contexts in which anyone might question the theory. In addition, the test should be efficient, requiring only minimal amounts of such resources as time and money.

One simple method to assess creativity is simply to ask people how creative they are, but such self-assessments inspire little trust. One reason is because people may lack insight into their own creative abilities. Another is that people...
may have reason to distort their assessments, either intentionally or unintentionally, especially if they believe the results will matter to them personally. A job interviewer who asks a candidate to assess her creativity would be foolish to accept such an assessment as valid without some other evidence. Even if there is no external reward associated with reporting higher levels of creativity, self-assessments are likely to be distorted simply because creativity matters to people, and higher self-assessments allow people to feel better about themselves. Because the degree of such self-deception will vary among individuals (and some might even err in the opposite direction, perhaps to conform to a low self image), one cannot just discount self-assessments a fixed percentage and thereby get a valid score. A test whose score that can be changed at whim is simply unlikely to be valid. Even if we could solve the problem of such intentional (or even unconscious) distortions, the problem of lack of insight would remain. Self-assessments are problematic because they are likely to include distortions of already suspect assessments.

So rather than asking how creative they are, one might instead ask people questions about themselves that are not directly linked to creativity. The adjective checklist, for example, asks participants to rate themselves on an extensive list of characteristics, some of which are believed to be associated with creativity (e.g., capable, clever, confident, egotistical, humorous, individualistic, informal, insightful, intelligent, interests wide, inventive, original, reflective, resourceful, self-confident, sexy, snobbish, and unconventional; Domino, 1994; Gough, 1960, 1979). Many self-report personality tests include creativity as a scale or subscale (e.g., the California Psychological Inventory). This approach has merit, but also significant drawbacks. There is certainly some transparency in the adjectives or personality test items, so participants may still be scoring themselves high to promote a particular image that might not be accurate. To the extent that the adjectives are not transparent, they may be linked to a particular theory of creativity (for example, being open to new experiences). This would mean that the validity of scores depends on the validity of the theory in question.

Some of the most widely used tests of creativity attempt to measure skills that are theorized to be components of creativity. Because these tests are not self-assessments and actually measure performance (albeit of a limited kind), these tests avoid some of the problems noted above. The most popular of these tests are divergent-thinking tests. Divergent-thinking abilities are believed by many (including us; see, e.g., Baer, 1993, 1997a) to be important creativity-relevant skills. However, calling a test of divergent thinking a “creativity” test (something suggested by the name of the most widely used divergent-thinking tests—the Torrance tests of creative thinking; Torrance, 1966, 1974, 1987, 1990, 1998)—may be a bit of a stretch. In the theory that first proposed divergent thinking (Guilford, 1956, 1967), divergent thinking was closely associated with creativity. Yet even Guilford maintained, “creative thinking cannot be equated with divergent thinking” (Torrance, 1988, p. 46).

A great deal of work has gone into assessing the validity of divergent-thinking tests, and although the tests are very frequently used and seem to correlate to some degree with other measures of creativity, it is probably fair to say that the jury is still out regarding the power and significance of divergent-thinking tests as measures of creativity (see, e.g., Baer, 1993, 1994c, 1994d; Cramond, 1994; Crockenberg, 1972; Kaufman & Baer, 2006a,b; Kim, 2006; Kogan, 1983; Torrance, 1969, 1972, 1981). So divergent-thinking tests may be related to creativity, but even if they are, calling these tests creativity tests is rather like calling a test of vocabulary an intelligence test. Divergent-thinking tests may be useful stand-ins for actual measures of creativity, but at best a limited one. Because they are so closely tied to one particular theory of creativity, divergent-thinking tests can hardly be used if one’s goal is to test rival theories of creativity.

The remote associates test (Mednick & Mednick, 1967) is another test that measures a skill theorized to be related to creativity. But like the Torrance tests of creative thinking, which measure a different skill rooted in a very different theory of creativity, even if the remote associates test is an accurate measure of the skill it claims to measure, it is nonetheless (a) assessing just one sub-skill that is at most a limited part of creativity and (b) too closely linked to a particular theory of creativity to allow its use in research that tries to make judgments about, or comparisons among, creativity theories. As a measure of creativity, therefore, it has serious limitations for the researcher.

So self-assessments are problematic because potential bias threatens their validity, and tests that measure skills that are theorized to be components of creativity, while possibly valid measures of those skills, are of somewhat limited use in creativity research because they depend for their validity as measures of creativity on the validity of specific theories of creativity that may themselves be participants of research. To avoid some of the problems of self-assessment and of theory-based assessment of creativity-relevant skills, one might simply ask other people to assess participants’ creativity, people who know them (or their work) well. However, there are some complications with this approach too.

Asking experts in a field to rate the creativity of the most famous practitioners in that field is fairly straightforward, because everyone in the field will know the work of the field’s leaders. Even more inclusive and exhaustive techniques
for probing expert opinion in a field are the kinds of citation counts and other objective measures that Simonton (1984, 1988, 1994, 1999, 2004, 2005, 2006) has employed so successfully. But these kinds of measures, while of extraordinary value for historiometric research, can be used only with participants who have produced a great volume of work. They also cannot be used in experimental research at all.

For most people who are not among the most eminent in their field, and for all research participants, regardless of their eminence, evaluation of creativity by other people who know the participants well would require different judges for each participant, making it impossible to compare ratings across participants. A creativity rating of 5 by rater X might be equivalent to a rating of 3 by rater Y and an 8 by rater Z, and there would be no way to calibrate such ratings to make them comparable.

The consensual assessment technique (CAT; see Baer, Kaufman, & Gentile, 2004; Conti, Coon, & Amabile, 1996) avoids most of these difficulties. In the CAT, participants are asked to create something, the creativity of which experts are then asked to evaluate. Poems, collages, and stories have been widely used in CAT studies, but the potential range of creative products that might work using the CAT is quite wide and includes musical compositions (Hickey, 2001) and drawings (Dollinger & Shafran, 2005). In the CAT, rather than try to measure some skill that is theoretically linked to creativity, it is the actual creativity of things participants have produced that is assessed, and it is assessed very directly.

The basic procedure when using the CAT is to give participants a prompt or instruction for creating some kind of product and then have 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). Experts in the area of children’s writing were then asked to rate the creativity of the stories on 1.0-to-5.0 scale. These expert judges were not asked to explain or defend their ratings in any way. They were 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. Inter-rater reliabilities among expert judges are quite good, typically in the .70-to-.90 range (Hennessey & Amabile, 1999; Runco, 1989).

The key question regarding the validity of any test is whether the test is measuring what it’s supposed to measure, and one of the great strengths of the CAT is how well it can answer this question. The CAT assesses the creativity of what might be called the garden-variety creative products (the poems, collages, etc.) of psychological studies the same way creativity is assessed at the genius level—by experts in that field. While it is true that experts do not always agree and expert opinion may change over time, at a given point in time there is no more objective or valid measure of the creativity of a work of art than the collective judgments of artists and art critics, just as there is no more valid measure of the creativity of a scientific theory than the collective opinions of scientists working in that field. If one wants to know how creative someone is as a poet, we do not look to psychological tests. We ask to see their poetry (or rather, we ask poets and poetry critics to look at their poetry).

CAT ratings of poems, stories, and collages have been shown to be valid measures of poetry-writing, story-writing, and collage-making creativity. Whether these measures also assess more general creativity-relevant skills is less certain (Amabile, 1983, 1996; Baer, 1993, 1994a, 1996, 2003; Conti et al., 1996), but for experimental studies designed to determine the impact of a wide variety of interventions, training, or experimental constraints, CAT ratings have been shown to work quite well. The CAT is not tied to any one theory of creativity, and so its validity does not rise or fall with one’s opinion of any particular theory. It focuses on creative performance, not creative thinking skills or other attributes that may be hypothesized to lead to creative performance. It is totally mute regarding most of the big questions in creativity research. For example, it can be used equally well by those who believe that creativity has a significant domain-transcending, general component (e.g., Amabile, 1983, 1996), those who argue for a more domain-specific understanding of creativity (e.g., Baer, 1994a, 1996, 2003), or even those who wish to separate domain-general and domain-specific variance in creativity (e.g., Baer, 1993; Conti et al., 1996). This would be impossible with any of the traditionally used divergent-thinking tests or with the remote associates test because such tests assume a high level of generality of creativity. CAT ratings are also generally stable across time (Baer, 1994b), but they respond quite well to real within-subject changes in motivation (e.g., Amabile, 1996) or skill (e.g., Baer, 1994a).

The CAT is not without limitations. Assembling panels of experts to rate those stories, poems, and collages is often expensive, although Kaufman, Gentile, and Baer (2005) found that gifted novices gave reliable ratings that were highly correlated with expert ratings. More pressing, however, writing stories and poems and creating collages are rather time-consuming. Might a task requiring only a few minutes, such as writing captions for pictures or cartoons, also
yield valid CAT ratings of creativity? If so, it would be a boon to creativity research. Potentially appropriate experts for evaluating the creativity of such captions would also be a much larger group, including college students (who are often both readily available and inexpensive to hire).

We can notice methodological limitations in CAT. In the conventional use of CAT there is an implicit assumption that there is only one facet in measuring creativity. In reporting the reliability of CAT judges have been seen to be the sole source of variation in measurement. With that practice it is hardly allowed to separate the influence of some other facets from that of judges in determining the precision of measurement. A facet is a universe of measurement conditions in test theory and it is like a factor in experimental design. The more factors are manipulated in an experiment, the broader perspective we can entertain in separating and understanding the sources of variation in the study. So is the effect of a facet in measuring a construct such as creativity. We are going to employ two facets in this study, which cannot be handled with the method of intraclass correlation or specifically inter-rater reliability in examining reliability of CAT.

Before we address study questions and method in this study, we would first like to clarify the concept of reliability. In reporting reliability, coefficient alpha or Cronbach’s alpha (Cronbach, 1951; Guttman, 1945) and some other concepts in classical test theory are most widely used. They provide information about consistency because they are based on correlations/covariances. The rating of two raters can be consistent “to the extent that an additive transformation serves to equate them” (McGraw & Wong, 1996, p. 33). When there is a need to determine relative standing of products, students’ performance, or job applicants, this type of reliability is relevant.

Whether the ratings agree between two raters, however, can also be described by agreement in absolute values. When the absolute values differ between raters, they are not viewed as in agreement, although they may be viewed as being consistent. McGraw and Wong (1996) provide a useful distinction between consistency and agreement. Since inter-rater reliability and inter-caption reliability deal with both consistency and agreement, we will provide a consistency index and an agreement index of the reliability of ratings in the present study. This approach should accommodate all readers, since some might be interested in relative standings or comparative judgments while others might be more interested in the degree of agreement in absolute values. When the number of facets is two or more, generalizability theory is needed to compute the degree of consistency and agreement in the ratings. Generalizability theory (Brennan, 2001) is an extension of classical test theory and intraclass correlation. This theory has a central focus on variance components of multiple facets in measurement. By referring to the size of the variance components, a researcher can explore optimal conditions of measurement by changing the number of levels in multiple facets and comparing the expected reliability corresponding to the change. In generalizability theory, reliability index for consistency is called the generalizability coefficient (“g-coefficient”) and that for agreement is called dependability index (“phi coefficient”).

The questions we tried to answer in this study are as follows:

1. Method

1.1. Design and participants

We employed two facets design to investigate reliability of ratings: four raters and twelve captions. Eighty-one undergraduate volunteers participated by writing captions. There were 43 males and 38 females. Sixty-five of the 81 participants were Caucasian, nine were Asian American, and seven were Hispanic.

Four graduate students (three in psychology and one in history) served as raters. Raters were selected randomly out of those willing to volunteer to complete the task. Raters were not familiar with the project and did not participate in the writing captions part.
Table 1
Reliability of ratings by each rater across captions

<table>
<thead>
<tr>
<th>Inter-caption reliabilitya</th>
<th>Raters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistencyb</td>
<td>.79</td>
<td>.50</td>
<td>.83</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>Agreement</td>
<td>.79</td>
<td>.50</td>
<td>.83</td>
<td>.75</td>
<td></td>
</tr>
</tbody>
</table>

a These coefficients were computed by the program GENOVA (Crick & Brennan, 1983).
b This coefficient is equal to α-coefficient.

Table 2
Reliability of ratings for each caption across raters

<table>
<thead>
<tr>
<th>Inter-rater reliabilitya</th>
<th>Captions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistencyb</td>
<td>.71</td>
<td>.72</td>
<td>.77</td>
<td>.78</td>
<td>.70</td>
<td>.64</td>
<td>.65</td>
<td>.78</td>
<td>.60</td>
<td>.72</td>
<td>.81</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>Agreement</td>
<td>.69</td>
<td>.70</td>
<td>.76</td>
<td>.74</td>
<td>.67</td>
<td>.61</td>
<td>.58</td>
<td>.78</td>
<td>.59</td>
<td>.69</td>
<td>.81</td>
<td>.76</td>
<td></td>
</tr>
</tbody>
</table>

a These coefficients were computed by the program GENOVA (Crick & Brennan, 1983).
b This coefficient is equal to α-coefficient.

1.2. Procedure

The caption-writing participants were taken to a psychology office and presented with a series of twelve black and white photographs. These photographs were selected from news archives for their ability to be open to multiple interpretations (see Kaufman, 2002, for more information).1 The students were then asked to write a few sentences for each photograph, with the instructions, “Please write a few sentences or a paragraph for each photograph. There are no set rules or expectations; just write whatever you feel should go with each photograph.”

The four raters scored all captions for all participants for creativity. Ratings were made independently and given a score from 1 to 7, with 1 being the lowest level of creativity and 7 representing the highest level of creativity. Raters were asked to assign creativity ratings based on their own personal definition of creativity; no additional guidance, descriptors, or material on creativity was provided.

2. Results

We offer the two types of reliability estimates via three alternative analyses: (a) inter-caption reliability of ratings by each rater across captions (the dimension of the data matrix is 81 × 12 for each rater) in Table 1, (b) inter-rater reliability of ratings for each caption across raters (the dimension of the data matrix is 81 × 4 for each caption) in Table 2, and (c) reliability of ratings across raters and across multiple captions (the dimension of data matrix is 81 × 4 × 12) at the bottom of Table 3.

The first two of these alternative analyses are related to the concept of intraclass correlation (McGraw & Wong, 1996; Shrout & Fleiss, 1979). In the theory of intraclass correlation, there is only one facet used in measuring the targets: captions in alternative (a), raters in alternative (b). In order to produce alternative (c) above, we applied generalizability theory and investigated the best method of designing creativity measurement within the framework of the CAT.

Table 3 shows the results of obtaining the variance component for each source of variation. We employed the program GENOVA (Crick & Brennan, 1983) to compute variance components (see Table 3) and the predicted reliability corresponding to different levels of the two measurement facets, raters and captions (see Table 4).

From Table 3, one can see that the variance component for people (.373) accounts for about 15.7% of the total variance in scores. This indicates that people are systematically different from each other to some extent in the creativity scores assigned to their captions. The variance components for captions, raters, interaction between people and raters, and the

1 Copies or descriptions of the photographs are available from the senior author.
interaction between captions and raters are not large. The interaction between people and captions is somewhat large, indicating that scores of people vary from caption to caption. The large residual variance (PCR) suggests (a) a large PCR interaction, (b) unknown sources of variability that were not measured in this study, or (c) both.

The information in Table 3 is useful in exploring the optimal conditions for CAT measurement in terms of numbers of raters and captions. The procedure for obtaining reliability estimates by increasing or decreasing the number of levels in raters and/or captions is much like the application of the Spearman–Brown prophesy formula in classical test theory (which Amabile, 1982, used in her initial work with the CAT). Although we employed four raters and twelve captions in the present study, the optimal number of conditions may be less than, equal to, or greater than the number of conditions that are employed.

Table 4 shows the expected reliability with different number of captions and raters. If our interest is limited to group-based inference instead of individuals, reliability of .6–.7 would be enough as we observe in Tables 2 and 4. However, if we draw inference on individuals, we would ask for higher reliability index such as .80 or more. If we want a reliability of .80 or more, Table 4 suggests that we would need to have at least thirteen captions and four raters for consistency in comparative judgment and at least fifteen captions and five raters for agreement among judges.

2.1. Interaction between captions and raters

Apart from the issue of reliability, we also wonder if there could be some captions that lend themselves more to creativity. We performed an analysis of variance with one between-subject factor (rater) and one within-subject factor (caption), with each participant’s creativity score as the dependent variable. The results are given in Table 5.

### Table 3

Variance components

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F^0</th>
<th>Variance component (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>People (P)</td>
<td>80</td>
<td>1808.13</td>
<td>22.60</td>
<td>4.79</td>
<td>.373 (15.7%)</td>
</tr>
<tr>
<td>Captions (C)</td>
<td>11</td>
<td>155/32</td>
<td>14.12</td>
<td>1.35</td>
<td>.011 (0.5%)</td>
</tr>
<tr>
<td>Raters (R)</td>
<td>3</td>
<td>287.58</td>
<td>95.86</td>
<td>9.74</td>
<td>.09 (3.7%)</td>
</tr>
<tr>
<td>PC</td>
<td>880</td>
<td>2863.07</td>
<td>3.25</td>
<td>2.76</td>
<td>.52 (21.8%)</td>
</tr>
<tr>
<td>PR</td>
<td>240</td>
<td>634.98</td>
<td>2.65</td>
<td>2.25</td>
<td>.122 (5.1%)</td>
</tr>
<tr>
<td>CR</td>
<td>33</td>
<td>276.07</td>
<td>8.37</td>
<td>7.10</td>
<td>.09 (3.7%)</td>
</tr>
<tr>
<td>PCR</td>
<td>2640</td>
<td>3109.62</td>
<td>1.18</td>
<td>1.18</td>
<td>2.381 (100%)</td>
</tr>
<tr>
<td>Total</td>
<td>3887</td>
<td>9134.78</td>
<td>2.381</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a For generalizability analyses, F-statistics are ignored.

b Consistency index of reliability provided in generalizability theory.

c Agreement index of reliability provided in generalizability theory.

### Table 4

Expected reliability for different number of raters and captions

<table>
<thead>
<tr>
<th>Number of captions</th>
<th>Number of raters</th>
<th>g-Coefficient</th>
<th>Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>.66</td>
<td>.63</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>.72</td>
<td>.68</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>.75</td>
<td>.72</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>.79</td>
<td>.75</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>.80</td>
<td>.76</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>.81</td>
<td>.78</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>.83</td>
<td>.79</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>.83</td>
<td>.80</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>.84</td>
<td>.81</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>.85</td>
<td>.82</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>.87</td>
<td>.84</td>
</tr>
</tbody>
</table>

a This design was actually used in the present study.
Table 5
Two-way ANOVA with rater and caption as independent variables and creativity score as dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
<th>Partial η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>295.72</td>
<td>3</td>
<td>98.57</td>
<td>12.88</td>
<td>.00</td>
<td>.11</td>
</tr>
<tr>
<td>S/rater</td>
<td>2426.61</td>
<td>317</td>
<td>7.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caption</td>
<td>161.48</td>
<td>11</td>
<td>14.68</td>
<td>8.77</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>Rater × caption</td>
<td>281.62</td>
<td>33</td>
<td>8.53</td>
<td>5.10</td>
<td>.00</td>
<td>.05</td>
</tr>
<tr>
<td>Caption × (S/rater)</td>
<td>5839.415</td>
<td>3487</td>
<td>1.675</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Two-way interaction between captions and raters.

Because there is a significant interaction between caption and raters, we have provided Fig. 1 to help clarify the nature of interaction.

Fig. 1 shows that the interaction is not ordinate, indicating that the order of scores assigned by raters varies across captions. Fig. 2 shows that caption 7 lends itself more to creativity. Judging from Fig. 1, however, this appears to be due to the serious interaction between rater 4 and caption 7. Rater 4 might have a valid reason for giving such high scores on caption 7, but if rater 4 were removed from the panel, Fig. 2 would become more flat.

2.2. Variance among raters: is it more by person or by caption?

Since we observe that ratings vary between raters, we are interested in knowing if they are more consistent or in agreement at a higher level for some people or some captions. In order to investigate this, we computed the variance among four raters in the cells of the $81 \times 12$ matrix. Since there is only one data point in the cells of this matrix, it

Fig. 2. Mean scores for each caption.
Table 6
One-way ANOVA with captions or people as independent variables and variance among raters as dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Captions</td>
<td>29.71</td>
<td>11</td>
<td>2.7</td>
<td>1.98</td>
<td>.03</td>
<td>.02</td>
</tr>
<tr>
<td>$S \times$ caption $^a$</td>
<td>1202.79</td>
<td>88</td>
<td>1.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1232.50</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>People</td>
<td>143.19</td>
<td>80</td>
<td>1.79</td>
<td>1.29</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>$S /$ people</td>
<td>1232.50</td>
<td>891</td>
<td>1.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1375.69</td>
<td>971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Subjects are 81 participants for each caption in this within-subject design.

$^b$ Subjects are twelve captions for each person (participant) in this between-subject design.

can be viewed as 81 observations for twelve captions (a within-subject factor) or as twelve observations for 81 people (a between-subject factor). Results of a one-way ANOVA with captions or people as the independent variable and variance among raters as the dependent variable are given in Table 6.

Variances are the inverse of consistency/agreement among raters; the more the variance, the less consistency (and vice versa). Table 6 shows significant main effects for captions and people. It indicates that there is at least one or more high mean value(s) of variance(s) for some participant(s) and for some caption(s). In terms of effect size expressed by partial $\eta^2$, such a tendency is more predominant for participants (people) than for captions. $\eta^2$ is interpreted with .01 as small and .06 as medium (Cohen, 1988). For captions, the size of $\eta^2$ is small, indicating that the difference between mean variances is minimal across captions. Since caption writing is a simple task that does not require a high level of expertise for rating creativity, it is to be expected that raters’ judgment would not vary much. However, $\eta^2$ for participants is close to medium size, indicating that raters show less consistency/agreement in judging some participants’ caption-writing creativity.

3. Discussion

High reliabilities were found in both inter-rater reliability and using generalizability theory, indicating that raters were able to agree on creativity across many different captions. With multiple captions and raters, we were able to obtain reliable measures of creativity. However, high reliability does not mean the captions are equal, of course, nor are the caption-writing participants equal. Rather, they vary, as shown in Tables 5 and 6. What we have demonstrated is that we can secure a desirable level of reliability in measuring creativity using a reasonable number of captions and raters. Multiple captions represent repeated opportunities for participants to show their creativity, allowing them to be observed a sufficient number of times. The use of multiple raters provides another opportunity for participants to be observed multiple times in another way—another facet, in terms of generalizability theory.

In summary, these results indicate that (a) captions appear to be a reliable way of measuring creativity; (b) raters agree on creativity across multiple prompts; (c) reliability in creativity ratings can also be demonstrated using generalizability theory; and (d) high levels of reliability (greater than or equal to .80), using either g-coefficient or phi, should be attainable using five raters who independently rate 15 captions for creativity. It should be noted that this is evidence for reliability of creativity ratings of individuals across several tasks, not reliability of ratings or single captions. Researchers would still need to ascertain that raters show agreement with each other.

Because having participants create 15 captions is not inordinately time-consuming and does not require special expertise, using captions as a measure of creativity has promise as a tool for creativity research, especially when domain effects (Baer, 1993) are not considered relevant. For example, in assessing the impact of varying motivational constraints, such as has been done in research by Amabile (1983, 1996) and Baer (1997b, 1998b) using poetry-writing, collage-making, and story-telling creativity, captions could be a very useful tool for measuring creative performance. It is worth noting that in Guilford’s work on the Aptitudes Research Project (ARP) at the University of Southern California, he included a test of Plot Titles, in which participants were asked to write titles for short-story plots (Wieder, 2005); and that Torrance (1990; Ball & Torrance, 1984; Torrance & Ball, 1984) includes a norm-referenced “Abstractness of Titles” subscale for the Torrance tests of creative thinking: figural (TTCT: F) in which test takers label drawings. This subscale is “based on the idea that creativity requires an abstraction of thought. It measures the
degree a title moves beyond concrete labeling of the pictures drawn” (Kim, 2006). There is also a criterion referenced measure called “expressiveness of titles” that is part of the TTCT-F scoring.

When using the CAT, one typically employs expert judges to assess creativity, judges who have appropriate expertise in the relevant domain. Thus when assessing the creativity of poetry, one must use poets, poetry critics, or perhaps teachers of poetry. Similarly, for evaluating the creativity of collages one would need artists, art teachers, or art critics. Because such experts are busy and generally need to be compensated for serving as judges (compensation in something more costly than extra credit of some kind), assembling such panels can be difficult, time-consuming, and expensive.

Who are the relevant judges of the creativity of captions? As far as we know, there are no Associations of Caption Writers, no degree programs in writing captions, and professors of caption writing. One could make the argument that some people do develop expertise in this area (such as advertising copywriters, newspaper editors, or certain types of humorists), but in each case, the expertise in caption-writing is a by-product of another, more far-ranging career. A journalist’s expertise in caption writing is much more incidental than a poet’s expertise in writing poetry. Certainly, there are some areas in which caption-writing expertise could be demonstrated (such as in the New Yorker’s caption-writing contests for cartoons). However, compared to most fields of creative production, caption writing does not have any specific, widely acknowledged experts.

One could argue that captions are somewhat unique among measures of creativity in the ease with which one can assemble an appropriate panel of expert judges. This would include the most readily available potential judges, college students. There is, however, a flip side to the ease with which one can find “experts” in an area like caption writing that is not a generally recognized field of creative endeavor. One could argue that the reason everyone is equally qualified to be an expert in this area is that there are simply no experts in this area and that judgments of creativity cannot, therefore, have the validity of expert judgments in fields like poetry or art. We think this concern is not one to be dismissed lightly.

At least we can suggest a post hoc method of detecting unqualified judges. Any judge/rater is expected to be consistent in measuring different captions. Consistency is defined by equal or similar pattern of judgment/ratings with other judges/raters. The raters might have different standards and thus assign different numbers from other raters to the first target of measurement. Once they start rating, their ratings should move in the same pattern: for example, when most raters’ scores ascend, any rater who assigns a descending score is not consistent. In situations where a rater is not consistent with other raters, she is interacting with the target of measurement (the captions in our case). For example, among our raters, rater 4 was not consistent with the other judges, as can be seen with captions 5 and 7 in Fig. 1. This is analogous to item bias in test theory. Item bias or differential item functioning is defined when an item is too difficult or too easy for a certain group of test takers—an interaction between item and group. A rater’s inconsistency can be called rater bias. A biased rater should be removed from the panel in the process of summing or averaging the scores for each participant before individuals’ scores are reported or arranged for other analysis.

We also by no means wish to suggest that someone who is creative in caption writing (as assessed by the CAT) is necessarily creative in other areas. This would need to be assessed (perhaps using correlational studies measuring CAT-assessed creativity in a variety of performance areas); such a claim could not be taken seriously until such studies have been conducted. Based on similar studies comparing creativity across a variety of performance domains, which have tended to produce quite low inter-domain correlations (Baer, 1993, 1994a, 1996, 1998a,b; Conti et al., 1996; Runco, 1987, 1989), we would in fact be surprised to find large correlations between assessed creativity in caption writing and creativity in very different domains, especially once variance attributable to g had been removed.

We are not, therefore, offering caption-writing creativity, as assessed by the CAT, as a general-purpose measure of creativity. We doubt that any single-domain test could make that claim (although perhaps assessing creativity using the CAT in a variety of domains, including caption writing, could produce a summed creativity-across-domains score of some kind—but such test-development and validation work remains to be undertaken and at this point is only speculative).

The CAT itself does not measure general-purpose, all-around creativity. It originated as a tool of social psychology for testing ideas such as task-motivational influences on creativity (Amabile, 1982, 1983, 1996). It is important in most creativity research that all participants can complete the tasks in some way and therefore that tasks used in CAT are fairly simple and do not need higher levels of expertise to solve them. Since our task of caption writing is simple, the mean variances in ratings across captions are minimal as seen in Table 6. This fits the goals and typical research use of the CAT.
The CAT was originally employed to draw inference about groups of people rather than to assess individual differences. It has been suggested, however, that the CAT can be applied to measure individual differences in creativity (Amabile, 1996). The effect size for participants is close to medium in Table 6. Medium effect size is generally interpreted as “meaningful and certainly worthy of study” (Keppel, 1991, p. 66). This suggests that researchers need to investigate for which participants the raters show high disagreement or variance in their ratings. This effort is especially required if the study using CAT is intended for drawing inference about individual differences. The participants might have to be measured again to provide a fair feedback on his/her creativity.

We propose, however, that for creativity research purposes, the assessment of caption-writing creativity using the CAT could be a very useful, low-cost, easy-to-use, and reliable tool. The information produced by this study has demonstrated its reliability, and has even given guidelines, as can be seen in Table 4, for the numbers of captions and judges to achieve whatever level of reliability researchers need for a particular study.

References


