Creativity Research Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653635

A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique

James C. Kaufman [a]; John Baer [b]; Jason C. Cole [c]; Janel D. Sexton* [d]
[a] Learning Research Institute California State University, San Bernardino
[b] Rider University,
[c] Consulting Measurement Group and Quality Metric,
[d] California State University, San Bernardino

Online Publication Date: 01 April 2008

To cite this Article: Kaufman, James C., Baer, John, Cole, Jason C. and Sexton*, Janel D. (2008) 'A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique', Creativity Research Journal, 20:2, 171 — 178

To link to this article: DOI: 10.1080/10400410802059929
URL: http://dx.doi.org/10.1080/10400410802059929

CREATIVITY RESEARCH JOURNAL, *20*(2), 171–178, 2008

ISSN: 1040-0419 print/1532-6934 online
DOI: 10.1080/10400410802059929

Routledge
Taylor & Francis Group

A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique

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

John Baer
Rider University

Jason C. Cole
Consulting Measurement Group and Quality Metric

Janel D. Sexton*
California State University, San Bernardino

The Consensual Assessment Technique (CAT) is one of the most highly regarded assessment tools in creativity, but it is often difficult and/or expensive to assemble the teams of experts required by the CAT. Some researchers have tried using nonexpert raters in their place, but the validity of replacing experts with nonexperts has not been adequately tested. Expert ($n = 10$) and nonexpert ($n = 106$) creativity ratings of 205 poems were compared and found to be quite different, making the simple replacement of experts by nonexpert raters suspect. Nonexpert raters' judgments of creativity were inconsistent (showing low interrater reliability) and did not match those of the expert raters. Implications are discussed, including the appropriate selection of expert raters for different kinds of creativity assessment.

One of the trickiest aspects of studying creativity is figuring out an appropriate way to measure the construct. One popular research technique for assessing creativity is the Consensual Assessment Technique (CAT; see Amabile, 1982, 1983, 1996; Baer, Kaufman, & Gentile, 2004). In the CAT, participants are asked to create something, the creativity of which experts are then asked to evaluate. This is essentially the same way that creativity and many other kinds of talent are assessed in the real world, whether it be for Nobel Prizes, Academy Awards, Fields Medals, *American Idol* competitions, Pulitzer Prizes, Grammy Awards, or National Science Foundation grant applications. In creativity assessment in the real world, it is common for panels of experts in a given domain to be asked to evaluate the creativity of some creative product or group of products (e.g., a work of art, a grant proposal, a theoretical model, a collection of poems, etc.). Some panels work and vote independently, and their assessments are averaged across the ratings of all experts, whereas on some panels the experts confer actively with one another to reach consensus. Despite such differences in technique, the model is consistent in this crucial respect: In evaluating creativity, only experts believed to know the domain well are asked to serve as judges and their combined evaluations are assumed to be the best possible assessment of

*Janel D. Sexton is now at Johns Hopkins University.

We thank Weihua Niu for assistance in setting up the Web site and collecting data; and James Bell, Mariah Bussey, Paul Dunton, and Kathleen Pelhaim-Odor for assistance with data entry.

Correspondence should be addressed to James C. Kaufman, Director, Learning Research Institute, California State University, Department of Psychology, 5500 University Parkway, San Bernardino, CA 92407. E-mail: jkaufman@csusb.edu

creativity in that domain at that time. It is recognized, of course, that later generations may revise earlier assessments, especially assessments of creativity at the highest, paradigm-shifting level (e.g., Csikszentmihalyi, 1996).

Amabile's (1982, 1983, 1996) pioneering work developing and validating the CAT for evaluating the creativity of diverse creative products has made possible a broad range of experimental studies of creativity. The CAT 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 interrater reliabilities that typically exceed .70, and often range as high as the .90s (e.g., Amabile, 1982, 1983, 1996; Baer, 1993, 1997, 1998; Kaufman, Baer, & Gentile, 2004; Runco, 1989).

Baer et al. (2004) were able to extend, significantly, the range of creative products that could be assessed using the CAT. The CAT, as originally developed by Amabile (1982, 1983, 1996), can be used only to compare parallel creative products—that is, ones created in response to the same assignments or prompts, such as the example in the previous paragraph. Baer et al. showed that the technique can also validly assess the creativity of nonparallel creative works—ones not created in response to the same prompts, but in response to very different assignments. This means that this procedure can now be used to compare such nonparallel assignments or prompts to determine, for example, which ones tend to produce higher levels of creative achievement. This extension of the CAT allows researchers to use existing data for creativity research that previously could not be done. There is, for example, a wealth of potential data in the various kinds of student works collected by the National Assessment of Educational Progress, as well as work produced and collected for other purposes. This recent extension of the CAT has, therefore, given a green light to researchers to make full use of this potential bounty of creative artifacts to evaluate diverse hypotheses regarding the factors that may be associated with, or tend to lead to, higher levels of creative performance.

This extension of Amabile's (1996) method does not free researchers from the need to assemble panels of expert judges, however. Such panels frequently include 10 or more experts, and this presents logistical challenges, especially when a single study involves assessing the creativity of a great many creative products. When a study involves creative tasks in several different domains, as is often required (e.g., in Baer, 1991, four separate panels were required to judge creative products in four different domains), the need for such panels of experts can become burdensome.

Some researchers have begun to use nonexpert judges (e.g., Niu & Sternberg, 2001). It is not clear how essential is it to use expert raters when using the CAT, nor exactly what kind of expertise is essential to be an expert rater in a given domain and for a particular research purpose. Can nonexpert raters reach consensus and provide appropriate judgments of creativity? This issue has been explored for nearly a century under the name *aesthetic judgment* (Cattell, Glascock, & Washburn, 1918). Many past investigations have found that expert-level judges consistently agree and have high interjudge reliabilities when judging artistic works (e.g., Child, 1962), even across different cultures (e.g., Child & Iwao, 1968; Haritos-Fatouros & Child, 1977; Rostan, Pariser, & Gruber, 2002). Some very initial work has been conducted on comparing novice and expert judgments. Runco, McCarthy, and Svenson (1994), for instance, found evidence that expert assessments in artwork may be harsher than peer or self assessments. However, between one and three experts were used for this study, which does not allow for a comparison of how well experts agree with each other versus how well novices agree with each other. Other studies have looked at aesthetic preferences or interests without looking at specific product ratings or agreement (e.g., Haritos-Fatouros & Child, 1977) or were unable to get sufficient expert agreement to conduct meaningful comparisons (e.g., Hickey, 2001; Runco et al., 1994).

One domain of creativity that is particularly relevant to be tested is creative writing. Reform efforts in school standards are showing a renewed interest in literature and creative writing (International Reading Association & National Council of Teachers of English, 1996). More than 50 colleges have decided to offer creative writing majors in the last 6 years (bringing the total to more than 300); this increase comes at a time when the number of English majors as a whole is decreasing (Bartlett, 2002). Yet only one study has examined novice versus expert evaluations in creative writing, and this study looked specifically at gifted high school students who were highly interested in the domain being rated (Kaufman, Gentile, & Baer, 2005). These gifted novices' ratings produced good interrater reliability and were significantly correlated with the creativity ratings of experts. However, gifted novices are not the same as nonexperts; they may fall somewhere in between the two groups. There is reason to believe that nonexpert judges, working independently (as required by the CAT), will achieve consensus in their ratings. Perkins (1981) noted that, for most people, "critical abilities are more advanced than productive abilities" (p. 128), and Johnson-Laird (1988) concurred, terming this the "central paradox of creativity" (p. 208). Whether

the critical abilities of nonexperts are sufficiently "advanced" (to use Perkins's term) to bring their evaluations in line with those of experts remains to be seen, however.

In her initial validation of study of the CAT, Amabile (1983) compared the creativity ratings of collages created by 22 girls (ages 7–11) of three groups of judges: psychologists in the Stanford psychology department, art teachers who happened to be taking a course at Stanford, and artists from the Stanford art department. The intergroup correlations ranged from .44 (between the artist–judges and the psychologist–judges; $p < .05$) to .65 (between the art teacher–judges and the artist–judges; $p < .01$). The psychologists lacked artistic expertise, but they did have a different type of expertise (i.e., understanding children). Therefore, these psychologists cannot be considered true nonexperts, nor were they randomly selected. Yet still there were differences in how well the judges with artistic expertise rated the creativity of the collages. Although the .44 correlation between the artist–judges and the psychologist–judges reached the .05 level of statistical significance, a correlation of only .44 hardly suggests that one group's judgments could simply be replaced by the others without effect (indeed, this relates to only 19.4% shared variance). In a separate study using a small sample of 20 collages created by undergraduates, Amabile (1983) found that 14 nonartists (undergraduate and graduate students not studying art and arguably, therefore, true novices) achieved considerable consensus (Spearman-Brown $\rho = .93$). But she did not compare these ratings to ratings of experts judging the same collages. Thus, the question of the validity of using nonexpert judges in research with the CAT is unknown, because high interrater reliabilities among nonexpert judges can assure only interrater reliability, and it is the use of panels of expert judges that has allowed the consensual assessment technique to claim that its ratings are also valid (for a more detailed defense of this validity claim, see Amabile, 1983; Baer, 1993).

If the creativity ratings of panels of nonexpert judges can be shown to match (or come close to matching) those of expert panels, however, then suitable panels of nonexperts judges could, in some research situations, replace more costly and difficult-to-assemble panels of experts, and this would facilitate future creativity research. If, on the other hand, the judgments of expert and nonexpert judges are found not to match, then this finding would provide guidance to researchers and editors regarding the use of nonexpert judges in creativity research and possibly help clarify what kinds of judges are most appropriate for a given creativity assessment task. The goals of this study were to determine whether nonexpert judges of creativity would also yield high levels of interrater reliability and, if so, to what degree creativity ratings of nonexpert judges match those of expert judges. It was hoped that this might also shed light on what kinds of expertise was most appropriate for particular assessment purposes.

METHODS

Step One: Writing the Poems to Be Evaluated

Participants. The participants who provided the writing samples consisted of 205 college students from two universities, one a private university from the northeast and the other a public university in the southwest. Participants took part in the study online for extra credit. The sample included 54 men and 151 women, with a mean age of 24.20 years ($SD = 8.73$ years). The demographic breakdown of the sample was as follows: 75 European Americans (56 women), 47 Asians (33 women), 47 Hispanics (25 women), 25 African Americans (16 women), and 21 with mixed backgrounds (19 women).

Procedure. The study was conducted online, where participants first read and signed a consent form, and then were given instructions for the task. Participants were given 10 min to write a SciFaiku poem. The SciFaiku (as the participants were told) is a form of poetry derived from haiku, a traditional Japanese poetry form composed of 3 lines of less than 17 syllables. The topic of the poem had to relate in some way to science fiction. See Appendix 1 for the complete SciFaiku instructions given to participants.

After completion of the study, all student writings were retrieved from the Web site and were identified only by the participant's numbers. All writings were printed in separate sheets with participant numbers on the top of each sheet. The participant SciFaikus were then prepared to be rated.

Step Two: Judging the Poems

Raters. There were two groups of raters: experts and novices.

Expert raters consisted of 10 poets who responded to a posting on a Web site for alumni of a poetry workshop. All raters were published poets, several with books of poetry to their credit and all with multiple publications in respected publications.

Novice raters consisted of 106 college students from California State University, San Bernardino, who participated in the study for course credit. Raters who participated in the first part of the study (writing the poems) were not included. The sample included 25 men and 81 women, with a mean age of 21.17 years ($SD = 6.21$

years). The demographic breakdown of the sample was as follows: 42 European Americans (30 women), 10 Asians (8 women), 37 Hispanics (31 women), 7 African Americans (6 women), and 10 from other ethnic backgrounds (8 women).

Rating procedures. Raters were given the poems in different, randomly assigned orders and asked to rate the poems for creativity on a 1 to 6 scale. To be consistent with the CAT methodology, the raters were asked to rate creativity of all poems, working independently. Neither group of raters was asked to explain or defend their ratings in any way. They were simply asked to use their own personal sense of what is creative in the domain of poetry to rate the creativity of the products in relation to one another.

Complete rater instructions can be found in Appendix 2.

Data Analysis

Prior to data analysis, missing data were addressed with modern techniques which have widely been found to be more proper than dropping participants with any missing data (Little & Rubin, 2002; Schafer & Graham, 2002). First, any raters who did not rate at least 75% of the writings, or any writers who were not rated by at least 75% of the raters, were omitted. This led to the removal of no data from the experts and led to the removal of 2 writers and 4 raters from the nonexpert database. Indeed, the expert database had no missingness at all. Next, the nonexpert database was subjected to multiple imputation data replacement. Multiple imputation uses an iterative regression-type approach in order to estimate each missing datum. Imputed values are generated, taking into account responses from the same participant on other correlated variables and responses to the same domain from similarly responding participants. Using such multiple imputation formulae, Rubin and Schenker (1991) have demonstrated that single imputation yields similar results to that of the more laborious multiple database process.

Consistency among the raters was evaluated with two measures of consistency: coefficient alpha (Cronbach, 1951) and a Spearman-Brown adjusted coefficient alpha (Bjorner, Damsgaard, Watt, & Groenvold, 1998). Coefficient alpha is a standard measure of internal consistency, and has been used in creativity research as a measure of interrater reliability (treating raters as items). For interpretative purpose, we used alpha levels of .90 or larger as excellent, .80 or larger as good, and .70 or higher as sufficient (Nunnally & Bernstein, 1994). As alpha is a point estimate, it is important to examine the standard error of the estimate as well. Because of the marked difference in the number of raters between the two groups, we also implemented a Spearman-Brown correction to alpha so that the two groups would be compared having equal number of raters (10), based on the consistency found among all of the raters in their group.

In order to compare significant difference between coefficient alphas, we used the standard error to calculate confidence internals (CIs). Estimates were compared at an alpha level of .05 for the difference. Nonoverlapping 95% CIs are not an indication of significance (Belia, Fiona, Williams, & Cumming, 2005). Instead, one needs to use 84% CIs to demonstrate a significant difference at a Type I error of .05 when the CIs do not overlap (Goldstein & Healey, 1995; Tyron, 2001). CIs for alpha were calculated based on the formula from Duhachek and Iacobucci (2004).

RESULTS

The novice raters rated a total of 204 poems. Their mean rating score was $M = 4.47$, $SD = 0.87$. The expert raters rated 205 poems and had a mean rating score of $M = 3.09$, $SD = 0.90$. See Table 1 for more details. An independent means t-test was calculated to compare mean ratings for the two groups, $t(407) = 15.64$, $p < .001$, (95% CI = 1.20–1.55). Given the respective means, the t-test indicated that expert raters rated the poems as less creative relative to novice raters. A Pearson correlation was assessed between the expert and the novice raters after determining that appropriate assumption checks were sufficient. Results revealed a significant relationship between the ratings that had an effect size between small and medium: r (202) = .216, $p = .003$, according to criteria from Cohen (1988).

Coefficient alpha for the expert raters was .832 (84% CI = .808–.856), which places this alpha in the lower range of a good alpha. Coefficient alpha for the nonexpert raters was .935 (95% CI = .926–.944, which places this alpha in the excellent alpha range. However, when the Spearman-Brown formula is applied to these alphas, the impact of the number of raters per group becomes quite clear. Standardizing the alpha to be set to 10 raters in each group (per the Spearman-Brown formula), we can see that alpha for the experts only dropped to .804 (95% CI = .771–.835), whereas alpha for the students dropped massively to .575 (95% CI = .562–.588).

TABLE 1
Descriptive Data for Novice and Expert Raters

	N	*Minimum*	*Maximum*	*Mean*	*SD*
Novice raters	204	2.56	6.88	4.47	.87
Expert raters	205	1.25	5.21	3.09	.90

DISCUSSION

Our discussion focuses on two issues. First, we consider what these results mean regarding the use of nonexperts to replace experts as judges when using the CAT. We then turn to the issue of what kinds of judges might be most appropriate for different kinds of CAT judging.

The CAT's validity is premised on the use of experts as judges. The results of this study clearly do not suggest that experts are easily replaceable. At least in the domain of poetry, experts and nonexperts showed differential levels of interrater agreement. Corrected alphas showed that expert judges agreed at a suitable level, whereas nonexpert judges did not. The nonexpert judges' average rating also did not closely match those of the experts' average rating, yielding a correlation of only .21. Despite the fact that this correlation was statistically significant, it is far too low (4.41% shared variance) to suggest that nonexpert ratings can be substituted for those of experts without changing the outcome. To make such a substitution would produce very different results.

The validity of the CAT is grounded in the fact that experts in a domain are the final arbiters of what is creative (or otherwise valued) in a domain. Coefficient alphas (or other measures of interrater reliability; see Amabile, 1983, 1996) are typically reported as evidence that the technique is working, but a high coefficient alpha only shows that the experts tended to agree in their independent judgments—that their ratings are not random, but rather reflect some shared variance. A high interrater reliability shows that there was something to rate, and that observable differences in creativity among the various products rated were found. Reliability is, of course, required before moving on to the larger question of validity. The validity of the ratings, however, cannot be assured merely by high interrater reliability alone. The judges' expertise provides a measure of face validity—it makes sense that experts in a domain could accurately assess performance in that domain. Yet without further study, it is impossible to ascertain whether reliable and appropriate judges provide construct validity. Expert judges may still use inappropriate standards, not understand the rating instructions, or have a biased agenda.

One limitation to this study may be that raters only gave a score for creativity, as opposed to several different scores related to creativity (such as originality). Although this methodology was used to be consistent with the larger body of research on the CAT, there may have been other differences. For example, Hekkert and van Wieringen (1996) examined aesthetic judgment of student art in experts and interested nonexperts. Although the two groups of raters did not score for creativity, they did score for originality, craftsmanship, and quality. The two groups' scores were significantly correlated for originality, but there was no relationship for craftsmanship and quality. A similar, more in-depth analysis may be warranted, although the high interrater reliabilities commonly attained suggest than any such factors must somehow be influencing the independent creativity raters of all, or almost all, expert judges.

The central finding, of experts and novices rating creative work in notably different ways, may be disheartening to those researchers looking at nonexpert raters as a quicker, cheaper substitute for expert ratings. If nonexpert judges could replicate the ratings of experts, it would make much creativity research easier, because nonexperts are far easier to find than experts in a domain. But, at least in the domain of poetry, that does not appear to be a reasonable expectation, and lacking evidence to the contrary, these results also suggest a need for caution in using nonexperts to rate other kinds of creative products.

What does it mean that the nonexperts and the experts rated the poems rather differently? One part of this result is probably not surprising: The experts gave the poems generally lower ratings. One reason for this finding is likely that they simply had higher standards when judging the poems. Because these expert judges are poets who generally read (and judge) much higher quality poems than those used in this study (which came from subjects who were not identified as poets or selected for poetry-writing skill), it is hardly unexpected that they would find the average creativity of the poems used in this study to be low. This possible discrepancy only addresses the differences in the *t*-tests, not the low correlations, however.

The poets were judging the poems, one might assume, based on internalized standards of what is creative in poetry. What standards might the novice judges have been using (because they, too, did not produce random results, and, in fact, achieved reasonable levels of agreement)? The novice judges also appeared to share a common metric, or set of standards, in their judgments, and although it was not the same as the metric of the experts, it was consistent. Are such novice judgments simply invalid? As direct substitutes for the ratings of experts, they clearly lack replicability; the two groups' ratings were not sufficiently correlated to allow one's ratings to substitute for the others'. But that does not make the ratings of the novices meaningless. Consider the awards such as the People's Choice Awards or the MTV Movie Awards. These awards poll large numbers of nonexperts, unlike the Grammys, the Emmys, and the Academy Awards, which rely on experts in their respective domains. The results of these awards are often quite different that the expert-based awards. In 2005, for

example, the MTV Movie Awards nominated several actors who have yet to be nominated for an Academy Award (or, perhaps, invited to come to the Academy Awards), such as Amanda Seyfried, Fred Armisen, and Tyler Perry.

The results of these and other popular awards are generally quite different from the standard award winners. Does this make the general public unfit to serve as judges? It depends on the goal of the evaluation. If one wants to know what TV viewers, moviegoers, and music listeners enjoy and prefer, then these popular awards may the best assessment. Indeed, if a studio wants a gauge of how well a movie will perform at the box office, an assessment from novices may be quite preferable to an assessment from traditional experts (and these concepts have been discussed in the marketing literature; see Holbrook, 1999; Holbrook, Lacher, & LaTour, 2006).

As Runco et al. (1994) argued, the best choice of experts may depend on the purpose of the assessment. If the goal is to find the most accurate assessment of creativity of products in a given domain, based on the current standards and values of that domain, then experts and other gatekeepers seem to be the group with the most face validity. Indeed, from a logical standpoint, they would seem to be most valid judges. In some fields, experts may be the only reasonable judges for most purposes. Asking the general public to evaluate the creativity of a new theory in nuclear engineering would make little sense.

Where does that leave the question of appropriate selection of judges for the CAT? If expert judges are considered to be the gold standard, then this study indicates that nonexperts and expert judges are not interchangeable. The question of which type of judge is preferable is still open for debate. There is clearly a need to choose appropriate judges for the particular creativity-judging task at hand—judges who know the domain, of course, but who, in addition to this domain knowledge, also have familiarity with the kinds of creative products typically produced by the kinds of subjects in the study. Experts in Big-C may not be the most appropriate judges of little-c. For example, judges for an elementary school science fair need to have some scientific expertise, but Nobel laureates in physics might be less appropriate raters than science teachers. Perhaps creative writing teachers with expertise in reading poems written by novices might be more appropriate than award-winning poets. The CAT still needs expert judges. But researchers need to be sure that the judges have the right kinds of expertise, which matches the kinds of products being assessed.

One intriguing possibility for future research is whether novice judges can be trained to be experts (or to provide ratings that correspond with expert ratings) through instruction or exposure. (Some early work has been conducted by Dollinger and Shafran, 2005.) It may well be that the vast discrepancy between the two sets of creativity ratings could be overcome with a basic training procedure. Clearly, such training would work better in domains requiring less domain-based knowledge; countless hours would need to be spent to instruct novices on how to become expert judges in creative theoretical physics. Another area to be explored is how dedicated novices who are interested in the subject matter differ from complete nonexperts. Early results from Myford (1989) indicate that theater buffs were completely in the middle between novices and experts in providing reliable ratings. How might the interaction play out with across-group agreement? The CAT has clearly opened many new doors for creativity research, but it appears that there is still have a great deal more to learn about how best to use it.

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. (1991). Generality of creativity across performance domains. *Creativity Research Journal*, *4*, 23–39.

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

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

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

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, *16*, 113–117.

Bartlett, T. (2002, March 15). Undergraduates heed the writer's muse. *Chronicle of Higher Education*, A39–45.

Belia, S., Fiona, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396.

Bjorner, J. B., Damsgaard, M. T., Watt, T., & Groenvold, M. (1998). Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *Journal of Clinical Epidemiology*, *51*, 1001–1011.

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

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

Child, I. L., & Iwao, S. (1968). Personality and esthetic sensitivity: Extension of findings to younger age and to different culture. *Journal of Personality and Social Psychology*, *8*, 308–312.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Csikszentmihalyi, M. (1996). *Creativity*. New York: HarperCollins.
Dollinger, S. J. & Shafran, M. (2005). Note on Consensual Assessment Technique in creativity research. *Perceptual and Motor Skills*, *100*, 592–598.
Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate estimate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*, 792–808.
Goldstein, H., & Healey, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society*, *158A*, 175–177.
Haritos-Fatouros, M., & Child, I. L. (1977). Transcultural similarity in personal significance of esthetic interests. *Journal of Cross-Cultural Psychology*, *8*, 285–298.
Hekkert, P., & Van Wieringen, P. C. W. (1996). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *American Journal of Psychology*, *109*, 389–407.
Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, *49*, 234–244.
Holbrook, M. B. (1999). Popular appeal versus expert judgments of motion pictures. *Journal of Consumer Research*, *26*, 144–155.
Holbrook, M. B., Lacher, K. T., & LaTour, M. S. (2006). Audience judgments as the potential missing link between expert judgments and audience appeal: An illustration based on musical recordings of "My Funny Valentine." *Journal of the Academy of Marketing Science*, *34*, 8–18.
International Reading Association & National Council of Teachers of English. (1996). Standards for the English/Language Arts: A Project of the International Reading Association and the National Council of Teachers of English. Newark, DE.
Johnson-Laird, P. N. (1988). Freedom and constraint in creativity. In R. J. Sternberg (Ed.), *The nature of creativity* (pp. 202–219). Cambridge: Cambridge University Press.
Kaufman, J. C., Baer, J., & Gentile, C. A., (2004). Differences in gender and ethnicity as measured by ratings of three writing tasks. *Journal of Creative Behavior*, *39*, 56–69.
Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, *49*, 260–265.
Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions consistency, creativity, and the consensual assessment technique: New evidence of validity. *Thinking Skills and Creativity*, *2*, 96–106.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.) Hoboken, NJ: Wiley.
Myford, C. M. (1989). *The nature of expertise in aesthetic judgment: Beyond inter-judge agreement*. Unpublished doctoral dissertation, University of Georgia.
Niu, W. H., & Sternberg, R. J. (2001). Cultural influences on artistic creativity and its evaluation. *International Journal of Psychology*, *36*, 225– 241.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.
Perkins, D. N. (1981). *The mind's best work*. Cambridge, MA: Harvard University Press.
Rostan, S. M., Pariser, D., & Gruber, H. E. (2002). A cross-cultural study of the development of artistic talent, creativity, and giftedness. *High Ability Studies*, *13*, 125–156.
Rubin, D. B., & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, *10*, 585–598.
Runco, M. A. (1989). The creativity of children's art. *Child Study Journal*, *19*, 177–190.
Runco, M. A., McCarthy, K. A., & Svenson, E. (1994). Judgments of the creativity of artwork from students and professional artists. *Journal of Psychology*, *128*, 23–31.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Tyron, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting nill hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

APPENDIX 1: SCIFAIKU POETRY INSTRUCTIONS

In the next page, you are asked to write a poem using the format called SciFaiku.

SciFaiku is a form of poetry derived from haiku, a traditional Japanese poetry form composed of three lines of less than 17 syllables. The topic is science fiction. It strives for a directness of expression and beauty in its simplicity. SciFaiku also frequently strives for insightful commentary on the human condition. Here is an example:

on blackhole's edge
indecision
drifts me in

You can also write more than one stanza, following the same rule of three lines of each. Here is another example:

hydroponics bay
a snail among stars
on the wide porthole glass.

mid-spring, anticipating
the imminent cloning
of humans

Bathing
her reptilian skin—
small bubbles on glossy green

In the space provided below, please write a SciFaiku poem, with a theme of science fiction. You can write anything you like, as far as your poem follows the rule of haiku (three lines of less than 17 syllables in one stanza). You should spend about 10 minutes on this, but please take your time.

APPENDIX 2: INSTRUCTIONS GIVEN TO RATERS

Please read through these poems twice. The first time, assign a Low, Medium, or High rating. The second time, assign a numerical rating between 1 to 6, with 1 being the least creative and 6 being the most creative. There should be a roughly even number of poems at each of the six levels, but the numbers needn't be exactly the same.

It is very important that you use the full 1–6 scale, however, and not assign almost all poems the same rating.

There is no need to explain or defend your ratings in any way; we ask only that you use your expert sense of which are more or less creative. Simply write the number on the paper (1, 2, 3, 4, 5, or 6—or, if you would find it helpful, any decimal from 1.00 to 6.00—but nothing below 1.00 or above 6.00, please).