Creativity has been proposed as a supplement to ability tests as a way to reduce bias, as a result of the typical lack of ethnic or gender differences. Yet, creativity is usually measured through a consensus of rater judgment. Could there be implicit biases against people of different ethnicities or gender? This study examined stories and poems written by 205 students and rated by 108 different students. Females wrote poems that were judged to be more creative; there were no significant differences by ethnicity across all raters. Among notable findings, European Americans and African Americans both preferred stories written by European Americans. Reasons for this and other findings are discussed.

When a person’s intelligence is measured, there are countless times when bias can enter the equation. However, the content of the response is rarely one of them. IQ subtests (e.g., Vocabulary, Digit Span, Picture Arrangement, Matrix Reasoning) have one correct answer. The essence of creating an IQ test comes in determining the relevant theory or structure, and then developing appropriate items.

Measuring creativity, however, is a different proposition altogether. The prompts used to elicit information can be quite generic (e.g., “Write a poem about snow”), but the content of the responses is incredibly subjective. One of the most commonly used ways to evaluate creative responses is the consensual assessment technique (CAT; see Amabile, 1982, 1996; Baer, Kaufman, & Gentile, 2004).

With the CAT, participants are asked to create something, the creativity of which experts are then asked to evaluate. 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. It has been used in many studies with a wide range of tasks (e.g., writing poems and stories, telling stories to go with pictures, creating collages and other artworks, creating mathematical word problems and puzzles) with both children and adults as subjects. In study after study, these expert ratings—conducted completely independent 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 .90 or higher (e.g., Amabile, 1982, 1996; Baer, 1993, 1997, 1998; Kaufman, Baer, & Gentile, 2004; Runco, 1989). Indeed, this technique is widely considered one of the best possible creativity assessments (Kaufman, Plucker, & Baer, 2008).

In this technique, the rater’s opinions are all-important. The school psychologist giving an IQ test can subtly influence the eventual score in a variety of ways, ranging from very basic ways (e.g., basic addition errors; see Slate & Hunnicutt, 1988) to more complicated ways (e.g., level of rapport that the psychologist establishes with the test taker; Gregory, 1999). Yet, the IQ test administrator is bound by the test in a way that the creativity test rater is not. When rating creativity, any biases (conscious or unconscious) can radically change the score that a person receives.

Any potential biases in creativity raters are especially important because creativity has been proposed as a possible way to counter test bias (Kaufman, 2005, 2006). Most studies that have examined ethnic or gender differences in creativity have used divergent thinking tests (for overviews, see Baer & Kaufman, 2008; Kaufman, 2006). Many studies have found few significant differences; while other studies have found females and European Americans with slight advantages on verbal measures, and males and African Americans with slight advantages on figural measures.

Those studies that have used the CAT have found little or no differences. Kaufman et al. (2004) studied poems, stories, and personal narratives written by African American and European American eighth-grade students. There were no differences across gender or ethnicity in creativity scores assigned by expert judges. Other studies have also used the CAT looking at gender (e.g., Amabile, 1996; Baer, 1993), finding no or little difference.

The results are less clear for Asian Americans, and only a small subset of Asian groups have been tested. Artwork produced by American college students was rated as more creative than art produced by Chinese students by both American and Chinese raters (Niu & Sternberg, 2001). Yet, a similar study that compared American and Chinese drawings of geometric shapes found that the two groups were rated similarly for creativity by both American and Chinese raters (Chen et al., 2002). There were no differences in rated artwork between Chinese and British school children, except for the higher ratings earned by Chinese children who attended a weekend art school (Cox,
Perara, & Fan, 1998). Another study found that Japanese children produced higher rated drawings than British children (Cox, Koyasu, Hiranuma, & Perara, 2001). Niu and Sternberg (2003) compared the creative performance between Asian Americans and non-Asian Americans, and found no differences between the two groups.

There have also been studies looking at how raters from different ethnicities or cultures differ (i.e., differences in aesthetic preference). Child and his colleagues (Child & Iwao, 1968; Haritos-Fatouros & Child, 1977; Iwao, Child, & Garcia, 1969) conducted a series of studies examining people’s aesthetic preferences. Their method was to ask participants to select an aesthetically superior picture from a pair. They also examined people’s aesthetic preferences in countries outside the United States (e.g., Japan, Israel). They did not find cultural differences in their studies. In addition, the earlier studies mentioned by Niu and Sternberg (2001) and Chen et al. (2002) did not find significant cultural differences across raters.

Kaufman (2006) asked 3,553 individuals (mostly high school and college students) to rate themselves in 56 different domains of creativity. Of the five factors derived from the 56 domains, males rated themselves higher than females on the science–analytic and sports factors; while females rated themselves higher on social–communication and visual–artistic. There were no differences on the verbal–artistic factor. African Americans were less likely to fall prone to gender stereotypes in creativity. In addition, African Americans and Native Americans tended to rate themselves as more creative than other ethnicities.

The interactions between the gender and ethnicity of the rater and the gender and ethnicity of the creator have not yet been studied. Based on past research, there should not be overall trends for raters to prefer creative work of one particular ethnicity or gender. Yet, do people unconsciously prefer creative work by their own ethnicity and gender? The goal of this study is to address this question.

Method

Step 1: Collecting the Creative Work

Participants

The participant writers consisted of 205 college students (54 males, 151 females) from two universities. Participants took part in the study for extra credit. The participants’ mean age was 24.2 years ($SD = 8.7$). The demographic breakdown of the sample was as follows: 75 European Americans (56 females), 47 Asian Americans (33 females), 37 Hispanic Americans (27
females), 25 African Americans (16 females), and 21 with mixed backgrounds (19 females).

Materials

The study used the CAT, as described earlier, to assess student creativity. We chose two written tasks, which fit the following major principles suggested by Amabile (1982, 1996): (a) the task should be open-ended enough to allow considerable flexibility; and (b) the task should not rely on any specialized skills so that everyone can finish it in a reasonable time, which can also display some degree of creativity. More specifically the two tasks were poem writing and story writing.

In writing a poem, the participants were instructed to write a SciFaiku poem in 10 min, following these instructions:

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 in²

In writing a story, the participants were instructed to select one of the following two titles—“2305” or “execution”—and then write a short story in 10 min.

Procedure

The study was conducted online. Participants first read and signed a consent form, read the online instructions, and completed their writing by typing words in the assigned boxes. After completion of the two writing tasks, the participants were asked to complete a brief demographic sheet.

After participants completed the study, all student writings were retrieved from the website, which were identified only by the participants’ numbers. All writings were printed in separate sheets with participant numbers on the top of each sheet. Two sets of writing were prepared—one for poems and one for

²This example was taken from the official website of SciFaiku at www.scifaiku.com
short stories. The writing was then ready for raters to evaluate participants’ creativity.

**Step 2: Rating the Creative Work**

*Raters and Rating Procedure*

Raters consisted of 108 college students (27 males, 81 females) of a public university in the Southwest, who participated in the study for course credit. Raters who participated in the first part of the study (i.e., writing poems) were not included. The raters’ mean age was 21.2 years ($SD = 6.2$). The demographic breakdown of the raters was as follows: 42 European Americans (30 females), 37 Hispanic Americans (31 females), 10 Asian Americans (6 females), 10 from other ethnic backgrounds (8 females), and 7 African Americans (6 females). Ethnic data were missing for 2 raters.

The participant raters were asked to read the SciFaikus and short stories and to assign the writings a score on a 10-point scale ranging from 1 (*lowest level of creativity*) to 10 (*highest level of creativity*). The raters were asked to assign creativity ratings based on their own personal definition of creativity; no additional guidance, descriptors, or material on creativity were provided. The raters did not meet or talk about their ratings with one another or with the experimenters. This format is consistent with the CAT, as developed by Amabile (1982, 1996) and as refined by others (Baer, 1993, 1998; Baer et al., 2004).

**Data Analysis**

Prior to data analysis, missing data were addressed according to modern techniques beyond the improper dropping of participants with missingness (Little & Rubin, 2002; Schafer & Graham, 2002). First, any rater or writer with greater than 25% missingness was dropped from the database, as even the best missing-data techniques have their limits. This led to the removal of 1 writer and 4 raters for the poems, and 3 writers for the stories.

Multiple imputation uses a regression-type approach to estimate each missing datum. Imputed values generated taking into account responses from the same participant on other correlated variables and responses to the same domain from similarly responding participants; an error in the estimation is also included to account for randomness around the perfect estimate that occurs in observed data. Using such multiple imputation formulas, Rubin and Schenker (1991) have demonstrated that single imputation yields...
virtually identical results to that of the more laborious multiple database process.

In order to answer the question of whether the raters had high levels of agreement in their ratings, intraclass correlations (ICC) were computed. The ICC is a statistic that measures the strength of relationship between a group of variables (e.g., raters’ scores). Different from a standard correlation, ICCs also take into account the absolute scores used by each rater, measuring not just similarity in correlation, but in score level as well (Shrout & Fleiss, 1979).

Additionally, internal consistency among raters was measured by the use of coefficient alpha for the poems and for the stories. For interpretative purposes, we evaluated an alpha of .90 or larger as excellent, .80 to .89 as good, and .70 to .79 as sufficient (Nunnally & Bernstein, 1994). As alpha is a point estimate, it was important to examine the standard error of the estimate as well. Therefore, we calculated the 95% confidence interval (CI) based on Duhachek and Iacobucci’s (2004) formula. In addition to the overall alpha levels of the raters, we also compared alphas between gender group and between European American and non-European American ethnic groups: once for stories and again for poems. Comparisons were conducted at an alpha level of .05. Nonoverlapping CIs are not an indication of significance (Belia, Fiona, Williams, & Cumming, 2005); therefore, the calculations were based on Fisher’s $z$-test transformations.

Results

**Intraclass Correlations**

With the modified data set, the average measure of ICC for poems was .92 (95% CI = .90–.93), and the individual ICC was .10 (95% CI = .08–.12). For stories, the average measure ICC was .91 (95% CI = .89–.92), and the individual ICC was .08 (95% CI = .07–.10). Differences between the average and individual ICCs reflect the tremendous augmentation to the ICC provided by having so many judges. Thus, the entire sample had great consistency as a whole, but any one judge compared to the total of judges was not particularly consistent.

**Comparing Interrater Reliability Coefficients Between Ethnic and Gender Groups**

Table 1 presents interrater reliability information via coefficient alpha estimates on poems and stories for each of four rater demographic
Table 1

Comparison of Interrater Reliability Between Genders and Between Ethnic Groups

<table>
<thead>
<tr>
<th></th>
<th>Poems</th>
<th>Short stories</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>-95% CI</td>
<td>+95% CI</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>.917</td>
<td>.901</td>
<td>.932</td>
</tr>
<tr>
<td>Males</td>
<td>.734</td>
<td>.681</td>
<td>.786</td>
</tr>
<tr>
<td>Ethnic group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Americans</td>
<td>.858</td>
<td>.831</td>
<td>.885</td>
</tr>
<tr>
<td>Non-European Americans</td>
<td>.896</td>
<td>.876</td>
<td>.912</td>
</tr>
</tbody>
</table>

*Note.* CI = confidence interval.
subgroups: females, males, European Americans, and non-European Americans. Please note that other specific ethnic groups were not used because the sample sizes were too small to allow for stability of the coefficient alpha statistic (Charter, 2003). For both poems and stories, female raters were significantly more consistent with one another than were male raters. Comparing between European American raters and all other raters reveals no significant differences in their consistency of ratings.

Comparing Overall Creativity Ratings by Rater Sex and Rater Ethnicity

We conducted two ANOVAs, with rater sex and rater ethnicity as independent variables and ratings of creativity for poems (first ANOVA) and stories (second ANOVA) as the dependent measures. A homogeneity of variance check was also conducted to ensure that there was not a violation of this assumption. Levene’s test of equality of error variance shows that the variances were comparable for each of the dependent variables ($p > .05$). Rater ethnicity did not predict overall ratings of creativity for poems, $F(4, 91) = 0.86, p = .49$; or for stories, $F(4, 96) = 1.54, p = .20$. Similarly, there were no differences in creativity ratings for poems, $F(4, 91) = 0.26, p = .61$; or stories, $F(4, 96) = 0.17, p = .68$, as a function of rater sex. There was also interaction of Sex ¥ Ethnicity found for ratings of poems, $F(4, 91) = 0.74, p = .57$; or stories, $F(4, 96) = 1.09, p = .38$.

This shows that raters’ ethnic group membership and sex did not predict the ratings they assigned to poems and stories. Because there were two outcomes measured for each participant (i.e., poem ratings and story ratings), it is best to consider an alpha adjustment here. However, given that all of the results were nonsignificant, adjusting the alpha lower would have had no impact on the results.

Comparing Creativity Ratings by Writer Sex, Ethnicity, and Age

Collapsing across all raters, the next question we turned to was whether raters as a group gave greater preference to certain ethnic groups or to one sex. To examine whether ratings varied by writer ethnicity, sex, and age, two separate multiple regressions were conducted: one for poems and one for stories.

For the analyses in the following sections, an adjustment for multiple outcomes was appropriate. We used an alpha adjustment based on the correlation between the two outcomes (Zhang, Quan, Ng, & Stepanavage, 1997). Given the family-wise alpha of .05 and the correlation between average story
rating and average poem rating of .414, each analysis will be compared at an adjusted alpha of .0335.

Poems

For poems, the overall model for the regression equation was significant, $F(3, 199) = 4.63, p = .004$. An examination of the beta weights shows that participant age and ethnicity were nonsignificant, but participant sex was significant ($\beta = .38), t(106) = 2.75, p = .006$. Females’ poems were rated as more creative than males’ poems.

Stories

For stories, the overall regression model was not significant. None of the beta weights for sex, ethnicity, or age were significant, either.

Do Raters Give Higher Ratings to Their Own Ethnic Group or to One Sex?

Separate contrasts were conducted for each ethnic group (except “other”) to answer the question of whether raters show preference for their own ethnic groups over other groups, and to compare the three major ethnic groups individually to European Americans. ANOVAs were conducted to examine sex differences by writer ethnicity and writer sex. Separate analyses were performed for poems and for stories.

African Americans

There were 7 African Americans raters and 25 African American writers. Poems. The two contrasts comparing African American writers to European Americans and to all other writers show no ethnic preference among the African American raters. A sex preference was found showing that female writers were preferred, $F(1, 201) = 10.30, p = .002$. The means for males and females were 4.42 ($SD = 1.32$) and 5.05 ($SD = 1.23$), respectively.

Stories. The contrast comparing African American writers to European American writers shows that European American writers were preferred, $t(195) = -2.63, p = .009$ (African American, $M = 5.08, SD = 1.25$; European American, $M = 5.72, SD = 0.84$). Similarly, the contrast comparing African American writers to all other ethnic groups was significant, $F(1, 195) = 2.36,$
\( p = .019 \), suggesting that African American raters preferred story writers who were not of their own ethnicity (African American, \( M = 5.08, SD = 1.25 \); others, \( M = 5.62, SD = 1.02 \)).

**Asian Americans**

There were 10 Asian American raters and 47 Asian American writers.

*Poems.* We performed two contrasts: one comparing ratings of Asian American writers to ratings of European American writers and another comparing ratings of Asian American writers to the ratings of all other ethnic groups collapsed together. For the first contrast, it was found that Asian American raters did not give higher ratings to Asian American writers when compared with European American writers, \( t(198) = 2.09, p = .04 \). For the second contrast, no ethnic preference among Asian raters was uncovered. A sex difference did emerge, however. An ANOVA reveals that among the Asian raters, female writers were preferred over males, \( F(1, 201) = 6.24, p = .01 \) (males, \( M = 4.14, SD = 1.31 \); females, \( M = 4.56, SD = 0.94 \)).

*Stories.* There were no differences in creativity ratings when comparing Asian writers with European American writers. There were also no differences when comparing Asian writers to all other non-Asian groups among Asian raters. Finally, higher ratings were not assigned to one sex over the other.

**Hispanic Americans**

There were 37 Hispanic American raters. In addition, there were 37 Hispanic American writers of poems and 34 Hispanic American writers of stories.

*Poems.* Contrasts compared ratings of Hispanic American writers with the writings of European Americans and to all non-Hispanic American ethnic groups. We found no ethnic preferences among Hispanic American raters. However, a sex difference was present. An ANOVA reveals that within the Hispanic American raters, female writers were preferred over males, \( F(1, 201) = 6.14, p = .014 \) (males, \( M = 4.14, SD = 0.94 \); females, \( M = 4.49, SD = 0.89 \)).

*Stories.* There were no differences in creativity ratings when comparing Hispanic American writers with European Americans or to all other writers, nor were higher ratings assigned to one sex.

**European Americans**

There were 43 European American raters and 74 European American writers.
Poems. A single contrast compared European Americans to all other groups. No ethnic preference was found among the European American raters. A sex preference was found showing that female writers were preferred, $F(1, 202) = 5.87, p = .016$ (males, $M = 4.25, SD = 0.78$; females, $M = 4.59, SD = 0.94$).

Stories. The contrast comparing European American writers to all other ethnic groups was significant, $F(1, 195) = -2.85, p = .005$ (European Americans, $M = 4.77, SD = 0.63$; others, $M = 4.48, SD = 0.75$). This suggests that European American raters preferred story writers of their own ethnicity.

Discussion

There are three notable findings that emerged from this study. Females wrote poems that were judged as more creative. Female raters were more consistent than were male raters. And European Americans and African Americans both preferred stories written by European Americans.

The first finding (i.e., females writing more creative poems) is not unexpected and is consistent with past findings that gender differences in creativity in verbal domains are either not present (as was found in stories) or favor females (see Baer & Kaufman, 2008; Kogan, 1974). Although past findings that have specifically looked at poetry have not found gender differences (Amabile, 1996; Baer, 1993; Kaufman et al., 2004), a number of studies that have used verbal measures of divergent thinking have found gender differences favoring females (Hines, 1990; Jaquish & Ripple, 1980; Kershner & Ledger, 1985; Kim & Michael, 1995; Rejskind, Rapagna, & Gold, 1992; Singh, 1979).

The second finding (i.e., females showing more consistency in their ratings) corresponds to the general finding that males show more variance in many different areas, from height to some cognitive abilities. As Pinker (2005) has summarized, this trend can result in more extremes. It is consistent that males would produce a wider variety of creativity assessments, whereas female assessments would converge more.

A more unexpected finding is the preference among African American raters for writers not of their own ethnicity. African Americans were the only group to rate the writings of another ethnic group higher than their own. There are several possible interpretations for this result. There is a past literature in social psychology that has held the view that people with lower status in society will have a more negative image of themselves. This originated with Cooley’s (1956) looking-glass self theory, which posits that people’s self-image is a reflection of their perception of how others view them. It follows from this perspective that the stigmatized would have a negative
self-concept, and thus be more self-disparaging. The notion of “self-hatred” among members of stigmatized groups has been written about and debated for decades (Cross, 1991; Lewin, 1941) and was bolstered in part by the Clark and Clark (1939) doll studies. Given that the raters were not aware of the writers’ ethnicity, their ratings can also be interpreted as an implicit measure of preference (Fazio & Olson, 2003).

An abundance of research, however, has found that self-hatred is not more likely to occur in stigmatized groups. In fact, many researchers have failed to find lower levels of self-esteem among members of stigmatized groups, and argue in contrast that members of stigmatized groups are able to protect their self-esteem (Crocker & Major, 1989; Cross, 1991). Recent views have dismissed the self-hatred theory, arguing that it is essentially racist (Baldwin, Brown, & Hopkins, 1991).

Another interpretation is simply that the African American preferences reflect what the dominant culture defines as creative. African Americans have historically not had a voice in defining what the culture deems to be creative; thus, their preferences may be largely influenced by societal standards of what is creative. Interestingly, Latinos and Asians did not show this pattern, so more research is needed to understand this phenomenon.

There are numerous areas for further study, many of which are related to limitations of the current study. The raters in the present study were college students, and not traditional “experts” (i.e., professional writers, teachers). It would be fascinating to compare possible bias (or lack thereof) in teachers and professionals. Indeed, given past research showing that experts and novices do not necessarily agree on what is creative (Kaufman, Baer, Cole, & Sexton, 2008), a further study that looks at potential bias in expert assessments would be particularly important.

In addition, the present study only examined creative writing. Given the evidence for the domain specificity of creativity (e.g., Kaufman & Baer, 2005), an exploration of creator–rater interactions in such areas as visual arts, science, or mathematics would be interesting and relevant. Writing is a very language-dependent task; in a domain without language (e.g., visual arts), cross-cultural comparisons would also be possible, such as those conducted by Child and his colleagues (Child & Iwao, 1968; Haritos-Fatouros & Child, 1977; Iwao et al., 1969). In addition, if further work is conducted using language-based creativity tasks, it would be useful to examine the content of the work from a linguistic point of view. Is there a noticeable difference in writing style or language usage in creative work across ethnicity and gender? Perhaps, despite the blind conditions of the study, raters were able to tell the gender or ethnicity of a particular piece.

In conclusion, these findings, which seem to indicate a possible implicit bias against African American creative fiction, are consistent with other
studies on implicit biases against African Americans (e.g., Dasgupta, McGhee, & Greenwald, 2000). They are especially concerning if creativity can, indeed, be a seen as a possible supplement to counter bias in measures of ability or intelligence (Kaufman, 2005, 2006). Given that this study found no differences among the different ethnicities on either poetry or stories across all raters, the reasons and specific instances of European American or African American bias against African American creative fiction should be investigated further.

References


