EMPIRICAL STUDIES OF THE ARTS, Vol. 26(2) 181-196, 2008

THE GENIUS PORTFOLIO: HOW DO POETS EARN THEIR CREATIVE REPUTATIONS FROM MULTIPLE PRODUCTS?

SCOTT BARRY KAUFMAN

ELISE M. CHRISTOPHER

Yale University, New Haven, Connecticut

JAMES C. KAUFMAN
Learning Research Institute
California State University at San Bernardino

ABSTRACT

How is creativity assessed across multiple products? What parameters influence the audience's overall impression of an artist's total body of creative work? This study examines this question in the domain of poetry, as poetry "gatekeepers" rated a series of five poems (all written by the same poet). The central question was what factors impacted the overall ratings of these poems; specifically, the following components were evaluated: average performance (i.e., typical work), maximum performance (i.e., best work), minimum performance (i.e., worst work), variability of performance (i.e., consistency), first performance, and last performance. The average, best, worst, and last poem in each set positively predicted the overall quality of the set. Variability (e.g., the standard deviation) did not make a significant prediction, suggesting that a body of artistic work may not be judged by the consistency of the set. These results suggest an overall stronger effect for ratings of individual items than for consistency when judging a set of creative works. Implications for aesthetic judgment are discussed.

© 2008, Baywood Publishing Co., Inc.
doi: 10.2190/EM.26.2.c
http://baywood.com

How is creativity assessed across multiple products? This dilemma is present across many fields. Scientists may offer multiple theories, and computer programmers may design many different types of software. Yet artistic creativity may offer the most commonly evaluated series of products, because art is readily available and able to be assessed—unlike the products of many fields. In many areas of the physical sciences, for example, only a fellow scientist would be aware of someone's full body of work. Yet millions of people may have seen several paintings by Johannes Vermeer, or read multiple Margaret Atwood novels.

When assessing the creativity of a set of performances, researchers have predominantly focused on either a single product or the average of a series of products (i.e., the typical performance). Such products are usually evaluated according to the Consensual Assessment Technique (CAT; see Amabile, 1982, 1983, 1996; Baer, Kaufman, & Gentile, 2004; Kaufman, Lee, Baer, & Lee, in press), where expert judges are asked to independently rate creative products based on their own conceptions of creativity. A great deal of past research (e.g., Amabile, 1983, 1996; Baer, 1993, 1998; Hennessey & Amabile, 1999; Kaufman, Baer, Cole, & Sexton, in press; Kaufman, Gentile, & Baer, 2005) has shown that experts agree at a strikingly high rate, with coefficient alphas that are usually above .70 and often above .90.

Most CAT studies have used the mean to serve as a proxy score of creativity. Typically, the products being assessed are created by schoolchildren or college students, not experts. Other research has studied aesthetic preferences, often by examining what components of artwork may capture someone's attention the most or be most pleasing (e.g., Cupchik, 1992; Smith & Smith, 2006).

However, what do experts use to evaluate creative artistic work when not doing so in the service of psychology experiments? The present study is designed to examine the parameters that play a role in quality judgments of a series of poems. This includes those works produced before the poet had reached her full potential, as well as ones produced at the height or end of his or her career. We make the distinction, however, that creative artists at the highest level have obtained enough expertise to be considered well-versed in their art form; we do not consider "expert" those performances produced while the artist was yet a novice. As Simonton (2003) explains, someone is a novice when even casual observers can discern one's lack of skills or knowledge in that domain.

The focus of this article is on poetry, as it reasonably meets the criteria for a creative domain. Poetry is a field in which "(a) the pressure for both originality and intelligibility is intense, (b) the products are invariably multidimensional and configurational, (c) the output rate for those products must be correspondingly low and (d) the reactions from the public, critics, and colleagues are mostly undifferentiated, inconsistent, and unstable" (Simonton, 2000, p. 287).

What are some candidate parameters? One possibility is that poets are judged by their most salient work, such as their best or worst poem. Various research results (see Gilbert, 2007, for a review) indeed demonstrate that individuals are apt

to remember the best and worst of times instead of the most likely of times. In a recent study, subway commuters who were waiting for their train were asked by researchers to imagine how they would feel if they missed their train (Morewedge, Gilbert, & Wilson, 2005). Commuters who were asked to remember any time they missed their train remembered episodes just as negative as those who were explicitly asked to remember the worst time they missed their train. The results suggest that when people think about their total experiences, often the single most inconvenient and frustrating episodes come to mind (Gilbert, 2007). It may be the case, then, that we make judgments of a set of work based upon the pieces that made us "feel" the most. If we attach an emotion to a piece of poetry, then we should able to remember that piece better and, further, to use it as an anchor for our judgments of other works by that artist.

Arguably, we implicitly apply this anchoring effect when forming overall impressions of a series of elements. Hayes (1983), in a study of ratings of academic *curriculum vitae,* found that what are perceived to be low-quality publications often hurt overall peer judgments of vita quality. Epstein (1985), in a follow-up letter, suggested that young authors should "publish what you please, where you want to or are able to publish it, and then be selective in listing publications on your vita" (p. 241). In other words, a larger output may actually be detrimental to how others perceive a body of work.

As a testament to the power of a single piece of work, consider the "one-hit wonder" phenomenon. Even though one-hit wonders produce only one lasting or significant piece of work in their particular domain, that work is played repeatedly or frequently cited, century after century, helping to ensure that the creator's name is remembered. Harper Lee, for example, received mass acclaim and a Pulitzer Prize (1961) for To Kill a Mockingbird, but has published virtually nothing since.

Likewise, a "rotten tomato" can seriously affect a poet's reputation. Kipling's "The White Man's Burden" (1899) is an example of a work that received such a bad response that it hurt his reputation and was even parodied in his day. This "rotten tomato effect" might function in various ways. There might be some threshold where a piece of work can be so bad that it is very hard for the artist to redeem his or her reputation. Alternatively, a single work might be so bad that it leads to a poorer judgment of the artist's overall performance; the anchor that drags the overall judgment down.

Placement in a set may also be an important parameter in forming an overall perception of an artist's work. Empirical research has indeed demonstrated that individuals often judge their total pleasure of an experience by its ending, whether they are thinking about their experience of pain (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993); child-rearing (Christensen-Szalanski, 1984), or marriage (Holmberg & Holmes, 1987). There is also substantial memory research that has found that people typically best remember the first pieces of information acquired (DiGirolamo & Hintzman, 1997; Miller, Westerman, & Lloyd, 2004) and the last pieces of information (Davelaar, Haarmann, Goshen-Gottstein, & Usher,

2006; Murdock, 1967). There is reason to suppose, then, that when sets of works are evaluated, effects of position (such as primacy and recency) may be coming into play.

Since it may be possible for creators to earn their reputation based on their best, worst, first, or last piece of work, this further suggests that it may be possible that the way an artist earns his or her reputation need not be confined solely to his or her typical performance, or even his or her consistency in performance. If one terrific piece of work out of a series of slightly above average works causes that artist to be perceived as terrific, then consistency is not the most important indicator of that artist's reputation. Similarly, if one terrible piece of work out of a series of good works causes that artist to be perceived as terrible, then the other good pieces did not seem to make much of a difference in the formation of overall impression.

In sum, experts may use salience, primacy, or recency effects when judging a body of artistic work, in addition or even in place of mentally taking the average of a set or judging the consistency of the set. Consistency may be paramount for building and maintaining expertise, but might not be enough (or could even be detrimental) to be considered creative at the master-level in an artistic domain (Kaufman & Kaufman, 2007; Simonton 2000).

The current study, therefore, is an attempt to investigate how these various factors influence a poet's reputation. We will investigate whether poets are judged by their (a) average performance (i.e., typical work), (b) maximum performance (i.e., best work), (c) minimum performance (i.e., worst work), (d) variability of performance (i.e., consistency of performance), (e) first performance, (f) last performance, or (g) some combination of the above.

METHOD

Participants

Participants consisted of poets and raters. There was some overlap between those who provided their poems, and those who rated poems. This did not pose a problem, as no rater was given their own poems to rate. Forty poets in total were included in the study. Thirty out of the 40 poets were published experts in the field with a wide range of levels of expertise—ranging from a graduate student in English to a published poet with over 50 years of experience writing poetry. These 30 poets provided a random sample of 5 of their poems that were published within a 1-year time period. The remaining 10 out of the 40 poets were Pulitzer Prize-winning poets taken from 1995-2005 (see Table 1). A random sample of 5 poems was taken from the book that won the poet his or her Pulitzer Prize.

In total, 40 sets of poems were assembled, with 5 poems (from a single poet) included in each set. Therefore, a total of 200 individual poems were rated in the current study. Each set of poems was conceptualized as a representative sample of

Table 1. Pulitzer Prize-Winning Poets Included in Sample

Franz Wright (2004)
Paul Muldoon (2003)
Carl Dennis (2002)
Stephen Dunn (2001)
C. K. Williams (2000)
Mark Strand (1999)
Charles Wright (1998)
Lisel Mueller (1997)
Jorie Graham (1996)
Philip Levine (1995)

Notes: 5 poems were randomly selected by each poet. Poems were taken directly from each poet's Pulitzer Prize-winning book. The year in parentheses refers to the year which each poet won the Pulitzer Prize.

that poet's total portfolio. These 40 sets of poems were mailed out to a sample of expert raters who all had expertise reviewing poetry for literary journals. Consequently, the raters in the current study could be conceived as "gate keepers" in the field of poetry (Csikszentmihalyi, 1996). Each rater rated 20 poems, split up into 4 sets. Therefore, the total sample consisted of 10 raters. Each set was read by one rater.

Rating Procedure

Each rater received a packet that consisted of 4 sets of poems. All poems remained anonymous, and raters were specifically asked after each poem whether they recognized the poem, or knew the poet. A sheet was included in the packet which explained to each rater the proper procedure for conducting the ratings. Each rater was instructed to score each *individual* poem along six dimensions (see Table 2). Use of the dimensions has empirical justification. A total of 13 expert judges in creative writing reached a consensus that these are crucial dimensions for the development of a poem (Baer et al., 2004; Gentile & Kaufman, 2002a, 2002b).

Each rater was also instructed to score his or her *overall impression* of the set according to the same criteria used for the individual ratings. A criteria sheet was included which explained the dimension, and the 1-5 rating scale used

Table 2. Rating Criteria

Subject Matter refers to the topic of the poem, and the ideas and feelings expressed in the poem. It represents the decisions writers make about what to say in a poem (the substance or material of the poem) as well as decisions they make about the overall approach they take to their material (the overall meaning or point of view of the poem).

e.g., Level 3**:** At this level, the treatment of the subject matter has some depth. Most of the ideas or themes presented in the poem are complete and/or developed. The writer has made specific decisions about what to say in the poem. In parts of the poem, it seems that the writer has begun to consider an overall meaning or approach to the subject matter.

Poetic Strategies refers to the use of techniques that are often associated with poetic and creative writing, such as imagery, metaphor, simile, personification, repetition, alliteration, onomatopoeia.

e.g., Level 3**:** At this level, there is a more overt use of strategies, but the strategies are used inconsistently. In one part of the poem the imagery is strong, the metaphor complete, but in the rest of the poem the imagery may be unclear or the metaphors confusing.

Poetic Devices refers to the use of conventions that are specific to poetry, such as rhyme, meter, rhythm, line breaks and layout.

e.g., Level 3**:** At this level, more poetic devices are used, but their use is still inconsistent or immature. The rhyming patterns tend to be "sing-songy;" the meter disjointed; and the rhythm and meter inconsistent. The writer has paid some attention to the layout of the poem, but the layout pattern is predictable. Often the use of poetic devices hinders rather than helps or does nothing to further the meaning of the poem.

Coherence and structure refers to the overall coherence of the poem: the degree to which the ideas in the poem flow smoothly or progress logically from one to another, and the structure or shape of the poem.

e.g., Level 3**:** At this level, the ideas are more focused and proceed in a logical or reasonable way. Poems at this level attempt to go somewhere but seem to get side-tracked along the way.

Effect on the Reader refers to the ways in which the poem affects the reader and the reader interacts with the poem.

e.g., Level 3**:** At this level, one part of the poem may strike a cord with the reader. The subject matter may have some depth (specificity, tension and/or emotion), but the use of poetic strategies and devices is inconsistent. The use of humor may make part of the poem interesting, but the rest of the poem may be disjointed, lacking coherence. Thus, the reader responds differentially, appreciating some parts but not others, not really interacting with the poem as a whole.

Imagination and Creativity refers to the novel use of the English language to convey an imaginative idea.

e.g., Level 3**:** At this level, the writing is about average in terms of imagination. Some familiar devices or imagery might be used, but there are also some creative new ones. Overall style may be average but the writer may have found striking ways to get across emotions, which makes the overall effect of the poem strong. This poem may leave the reader wishing for a little more, since it shows potential.

for each dimension. Raters were instructed to view each set as independent of the others. Half of the raters were instructed to rate their *overall* impression of each set first, whereas the other half rated each *individual* poem first. This counterbalancing was employed to ensure that the results of the study are based on an overall impression formation, and not a reliance on the individual scores in each set.

Measures

The current study was conducted twice, on two independent groups of raters. This was done so that we could assess how well the results from the first group of raters would replicate to an entirely new group of raters. For both group of raters, all six rating dimensions (Subject Matter, Poetic Strategies, Poetic Devices, Coherence and Structure, Effect on the Reader, and Imagination and Creativity) significantly correlated with each other. In rater group 1, the lowest intercorrelation between the dimensions was .77 ($N = 40$, $p < .01$), and for rater group 2, the lowest correlation was .89 ($N = 40$, $p < .01$). Since the correlations between the dimensions were so high, the averages of all the dimensions were combined to form a single quality score for each poem in a set.

For each set of 5 poems, we characterized the data in 6 ways, each characterization corresponding to one of the 6 factors as discussed in the introduction. Specifically, for each set of poems, we computed:

Mean score ("typical" performance)
Standard Deviation (consistency of performance)
Highest rated poem in each set ("Best" poem)
Lowest rated poem in each set ("Worst" poem)
First poem in each set
Last poem in each set

Controls

Four control variables were introduced. They were: gender of rater, gender of poet, order of presentation, and Pulitzer Prize winner. Gender of rater and gender of poet were both dummy coded "1" for male, and "2" for female. Order of presentation was dummy coded "1" if the rater first rated each individual poem in the set, and "2" if the rater first rated the overall quality of the set. This was introduced to control for order effects. Pulitzer Prize winner was dummy coded "1" if the poet that was being rated was *not* a Prize winner, and "2" if the poet *was* a Pulitzer prize winner. This control was introduced to ensure that the results would be generalizable across varying levels of expertise, and not confined solely to perceptions of poetry at the highest level.

Even though we controlled for these four variables, it is still of considerable psychological interest how each of these controls influenced overall ratings.

As for gender of rater, overall ratings were higher for males ($M = 4.08$, $N = 4$) than females ($M = 3.11$, $N = 36$) in rater group 1, and this difference was statistically significant ($p = .05$). In rater group 2, overall ratings were numerically higher for females ($M = 3.17$, $N = 28$) than males ($M = 2.81$, $N = 12$), although this difference wasn't statistically significant. Therefore, the gender of the rater didn't have a consistent influence on overall ratings across the two groups of raters.

As for gender of poet, overall ratings were numerically higher for male poets (Rater Group 1: $M = 3.38$, $N = 14$; Rater Group 2: $M = 3.89$, $N = 14$) than female poets (Rater Group 1: $M = 3.11$, $N = 26$; Rater Group 2: $M = 2.61$, $N = 26$), although this difference was only statistically significant in rater group 2 ($p < .01$).

As for presentation order, overall ratings were numerically higher for those who first rated each poem in the set individually before forming an overall impression (Rater Group 1: $M = 3.44$, $N = 20$; Rater Group 2: $M = 3.22$, $N = 20$) than participants who formed an overall impression of the set before rating each individual poem (Rater Group 1: $M = 2.96$, $N = 20$; Rater Group 2: $M = 2.90$, $N = 20$). This finding was not statistically significant for either rater group.

As for expertise of poet, both groups of raters rated the Pulitzer Prize-winning poems (Rater Group 1: $M = 3.67$, $N = 10$; Rater Group 2: $M = 3.78$, $N = 10$) more highly than the rest of the poems (Rater Group 1: $M = 3.05$, $N = 30$; Rater Group 2: $M = 2.82$, $N = 30$). Although this difference was only statistically significant in rater group 2 ($p < .05$), the difference approaches significance in rater group 1 ($p = .07$). Therefore, out of the four controls, the expertise of the poet demonstrated the most consistent effect across both groups of raters.

This finding provides external validity to the study, and suggests that raters were taking their jobs seriously. If the raters were filling out the ratings at random, one would not expect the Pulitzer Prize-winning sets to be systematically rated higher than the rest of the poem sets, as all of the sets were anonymous and therefore the raters had no prior knowledge that any of the sets were from Pulitzer Prize-winning poets, let alone who the authors were. (Note: as mentioned above, raters were asked for each poem whether they recognized either the poem or recalled the poet. Only one rater for one poem was able to accurately recall the poet, and most of the time this section was left blank.)

RESULTS

A summary of all of the results are listed in Table 3. First, we entered all six of our predictor variables (Mean of set, Standard deviation of set, Best poem in set, Worst poem in set, First poem in set, Last poem in set) together into a stepwise regression model, to assess which measures independently predict the overall quality of each set. For both groups of raters, controlling for the gender of rater, gender of poet, order of presentation, and Pulitzer Prize-winning status, and

Table 3. Summary of Regression Results for Both Groups of Raters (*df* = 39) Controlling for Order of Presentation, Gender of Rater and Poet, and Pulitzer Prize

	Rater Group 1		Rater Group 2	
	β	*p*	β	*p*
Predictors entered stepwise (Mean, Standard Deviation, Best, Worst, First, Last)				
Mean	.50	.01	.96	.001
Best	.54	.01	—	—
First	–.25	.04	—	—
Total R^2 =	.74		.92	
Predictors entered stepwise (Best, Worst, First, Last)				
Best	.54	.001	.40	.001
Last	.37	.01	.34	.001
Worst	—	—	.29	.01
Total R^2 =	.73		.90	

Note: β refers to the standardized regression coefficient.

controlling for the other four measures (Best, Worst, First, and Last), the Mean of each set significantly predicted the overall quality of each set (Rater Group 1: β = .50, *df* = 39, *p* = .01; Rater Group 2: β = .96, *df* = 39, *p* = .001). Additionally, for rater group 1 (but not for rater group 2), the Best poem in each set was a significant positive predictor above and beyond the mean (β = .54, *df* = 39, *p* = .01), and the First item in each set was a significant negative predictor above and beyond the Mean (β = -.25, *df* = 39, *p* = .04). The Mean, Best, and First measures explained 74% of the total variance in overall ratings for rater group 1, and for rater group 2, the Mean explained 92% of the total variance in overall ratings.

Since the Best, Worst, First, and Last measures share a significant portion of variance with the Mean (in effect these four measures are subsets of the Mean), we excluded the Mean from the regression model, and ran a stepwise regression using Best, Worst, First, and Last as the predictor variables. For this analysis, we also excluded the standard deviation, since it wasn't a significant predictor in the first analysis. For the second group of raters, Best (β = .40, *df* = 39, *p* = .001), Last (β = .34, *df* = 39, *p* = .001), and Worst (β = .29, *df* = 39, *p* = .01), made *independent predictions* on overall poem set quality. For the first group of raters, Best (β = .54, *df* = 39, *p* = .001) and Last (β = .37, *df* = 39, *p* = .01) made an independent prediction on overall poem set quality. For rater group 1, the Best and Last measures explained 73% of the total variance in overall ratings, and for

rater group 2, the Best, Last, and Worst measures explained 90% of the total variance in overall ratings.

It is possible that these results could have been influenced by habituation effects. Repetition of similar things invariably causes a decrease in preference (Berlyne, 1971). To rule out mere repetition effects, we looked at trends across all the poems a rater judged rather than confining the analysis only to sets by a specific poet. Figure 1 shows the results for both rater group 1 and 2, separating those who conducted individual ratings first, and those who conducted overall ratings first. In no case is there a decrease in ratings. If anything, there appears to be a trend toward an increase in ratings. Therefore, the results of the study don't seem to be affected by repetition effects.

DISCUSSION

For both groups of raters, the Best and the Last poem in the set made *independent* predictions on the overall quality of the set when the mean and standard deviation were excluded from the regression model. In rater group 1, both the Best and First poem in the set predicted overall ratings above and beyond the Mean (although the First poem was a negative predictor). Furthermore, the Worst poem in the set made a significant independent *positive* prediction on overall ratings in rater group 2, once the Mean and Standard Deviation were excluded from the regression model.

That the Best and First poems in the set predicted overall ratings suggests that both salience and position effects come into play in the perception of overall performance. The significance of the Worst poem in rater group 2 suggests that if a poet's worst poem is still better than other poets' worst poems, he or she will receive an overall higher rating.

That the Standard Deviation was not a significant predictor of the overall ratings in either group of raters suggests that experts in creative fields such as poetry may indeed not be earning their reputation due to their consistency of performance (Simonton, 2003). Therefore, the way a professional earns his creative reputation may depend on his domain. Consistency may be more important in domains that emphasize the consistent application of expertise, such as medicine and architecture. A neurosurgeon earns his reputation by consistently saving lives, an architect earns his reputation by designing houses that do not fall apart, a professional cellist in an orchestra (as opposed to a solo cellist) earns his reputation by having consistently technical auditions and performances, and a basketball player needs to consistently make foul shots and gain rebounds. In these domains, one or two standout performances may not influence the professional's reputation, if the rest of his performances are not consistent.

In support of this view, Simonton (2003) characterizes the life of the archetypal creator as a "sequence of hits and misses, of success and failures. . . . This hit-or-miss feature of the creative career contrasts immensely with what is

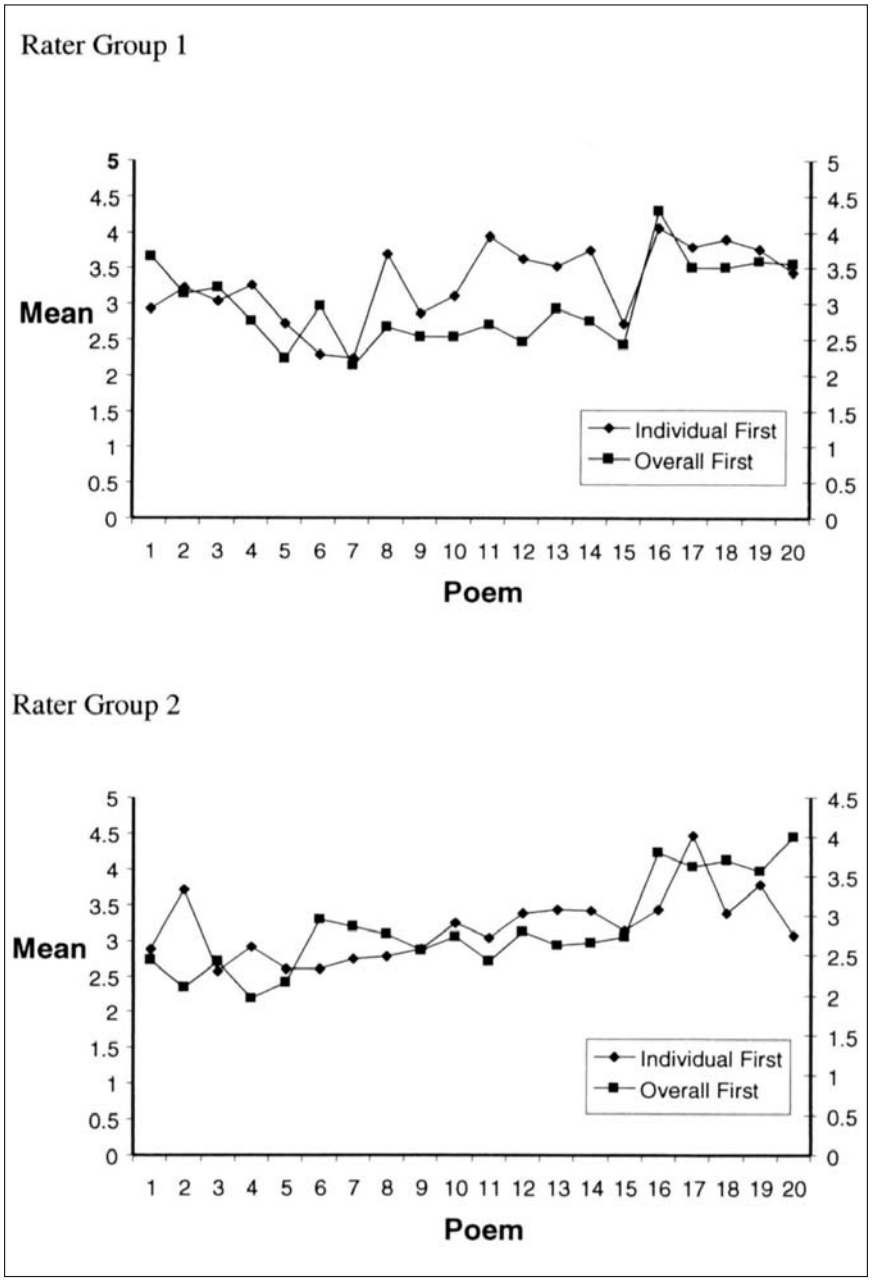

Figure 1. Overall mean ratings of each poem, collapsing over poem set.

observed in those achievement domains where the importance of expertise is unquestionable" (p. 230). Simonton argues that there may be adaptive gains to a more erratic style. Creators who "mix it up" or even experiment with different media, at the risk of producing a rotten tomato or two, may be increasing their chances of working something new into their repertoires, thereby ensuring their creative vivacity. This tendency may be especially true of creation in the arts; consider the consistent finding (e.g., Feist, 1999) that conscientiousness is generally positively related to being a creative scientist but negatively related to being a creative artist.

The results from the present study seem to point to the usefulness of several effects when making judgments of overall set quality. The "rotten tomato" effect and the "one-hit wonder" effect both emphasize the importance of one individual work's quality when compared with the rest of the set. The recency effect seems to place weight on an item's position, but this effect belies the main fact that the last item in a set is still an *individual* item, and in the current study affected the overall set ratings in similar fashion to the best and worst items. This result is interesting precisely because of the common tendency to focus upon the overall consistency of an artist in subjective discussions of performance quality. By giving expert poets the explicit instruction to rate poems both individually and as members of sets, we were able to show that the tendency of the expert poet's overall impression of the set to be influenced by the consistency of the set is often overridden by individual ratings.

In real world evaluations of a body of creative work, the effect of the best or worst work may be related to the overall quantity of the set. Poets who produce a distinctive masterpiece may have larger bodies of work and therefore readers have more to draw from when making judgments of greatness. W. H. Auden noted, "the chances are that, in the course of their lifetime, the major poet will write more bad poems than the minor" (quoted in Bennett, 1980, p. 15). There is empirical support for this statement. Simonton (1977) divided the lives of 10 eminent composers into 5-year periods, and measured each composer's productivity based on both their works and their themes. Simonton then found that the composers who wrote the most music also wrote the best music. The most fertile time periods in terms of production were also marked by the best work. Simonton (1985) also found this same effect with psychologists. It is interesting to note that this connection between larger quantity and higher quality would seem to indicate that the tendency of academic evaluators to downgrade longer vitas with lesser works (e.g., Hayes, 1983; Epstein, 1985) is quite misguided.

Conversely, an artist with a small body of work, among which one is a "rotten tomato," may come to be seen as a hack. Support for this idea comes from the tendency for creative performers to be forgiven for bad work if they produce enough good work. An example is Wordsworth, whose later poetry is generally agreed to be greatly diminished in quality, a point that is typically overlooked in evaluating his place in the canon because his early poetry is so very good

(Duemer, 1991). The current study does not address this issue, since quantity of poems was not manipulated. An interesting future line of research would be to manipulate quantity and see the effects of various parameters (such as those used in the current study) on overall perceptions.

It is interesting that all of the dimensions (see Table 2) used to rate each poem were so highly correlated with each other. This suggests that poetic quality may be unidimensional, and future studies on poetry may be able to abandon the notion of separate criteria. Future studies should see if this unidimensionality replicates.

It should also be noted that since we obtained a random sample of poems for each poet, "First," "Last," "Best," and "Worst" do not necessarily correspond to initial, concluding, best, and worst poems in that poet's career, but only to that particular, randomly-selected set of 5 poems. The benefit of using a random sample is that we were able to maintain objectivity in the selection of poems to include in the sample. A disadvantage is that the current study cannot speak to the issue of the poet's creative development over time, an issue that is certainly central to real poetry portfolios. In the real world, a poet's developmental arc may be influenced by life events and maturity. Indeed, Simonton found that poets peak markedly earlier than other writers (Simonton, 1975, 1997) and poets produce twice as much of their lifetime output in their twenties as novelists do (Simonton, 1984). A study of a poet's complete works would likely reflect such patterns (and would likely be different from a similar study of a novelist's complete works).

Another disadvantage is that it is very unlikely that the very best and very worst poems of that poet's entire career were included in any of our sets, and the first and last were certainly not included. Therefore, the chances are high that most poems selected will be mediocre, reducing the variance (and the standard deviation). Future studies should see if the results of the current study replicate when the actual best and worst products (i.e., most and least anthologized) are included in the sample.

Of further note is that the current study may generalize more to high art than popular art. Further studies should attempt to use different art forms, not only to generalize the current results beyond poetry, but also to generalize beyond high art. In high art such as poetry, critics have already screened the work for the average reader in various ways. For instance, few people who are not English majors have read the complete works of any minor poet. Yet many people have read the complete works of a popular thriller writer such as Harlan Coben, or have listened to the complete works of popular singers such as Billy Joel.

The results of this study have practical implications for artists. If someone is compiling a book of poetry or a music CD, or is arranging paintings in a museum, the present findings suggest that an excellent overall impression might best be cultivated through the use of individual works. The artwork in the "middle" will have the least effect on a reader, viewer, or listener—the best material should be placed late to have maximum impact.

It is also interesting to note the relative lack of importance of the First work. Despite the idea of "first impressions," the quality of the First poem had comparatively little importance in the overall evaluation of a poet's creativity (and may even have a detrimental effect). It may be very tempting to present your Best poem or song First to demonstrate your potential, but this study indicates that this is not the best strategy.

Most models of aesthetic preferences focus on the appreciation, emotional impact, or enjoyment of one particular piece. Leder, Belke, Oeberst, and Augustin (2004), for example, present an extensive model of aesthetic appreciation and aesthetic judgments in which they integrate such concepts as perceptual analysis, implicit memory, the context of the artwork, and several related cognitive processes. Yet the model is still primarily focused on the experience of one piece of art. Since much creative work is experienced in sets of works rather than individually, it seems only sensible to study the judgments of portfolios rather than merely single pieces. Of course, we must not ignore the interaction between the judgment of sets and of the individuals. There are many properties of set judgment that are not well understood, as the results of the present study imply. For example, the idea of a quality threshold, below or above which an individual judgment begins to shape the cumulative judgment, bears further study. We hope that more research (and, eventually, theoretical frameworks) will look at multiple pieces and their effect on how one perceives an artist's overall work.

ACKNOWLEDGMENTS

The authors would like to thank the following Poets and Raters for their valuable time and effort: Dick Allen, Madeline Artenberg, Caitlin Barrett, Norma Bernstock, Marie Boroff, Susan A. Clark, Cortney Davis, Audrey Fitting, Joan Foran, Anthony Fusco, Midge Goldberg, Marj Hahne, Sean Harrigan, Norbert Hirschorn, Suzy Lamson, Leslie McGrath, Sheila Murphy, Jennifer Nelson, Stella Padnos, Nii Parkes, Norah Pollard, Anna Reisman, Charles Rafferty, Bessy Reyna, Robin Sampson, Iris Schwartz, Lisa Siedlarz, Sonya Taaffe, Chris Tusa, Faith Vicinanza, Chocolate Waters, Kelley White, Liza Wolsky. The authors would also like to thank John Baer for helpful suggestions on an earlier draft, Colin DeYoung for assistance with the data analysis, and Robert J. Sternberg for input at the design stage of the study. Finally, the authors would like to thank Colin Martindale and two anonymous reviewers for their helpful suggestions.

REFERENCES

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

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

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

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

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

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*(1), 113-117.

Bennett, W. (1980). Providing for posterity. *Harvard Magazine, 82*(3), 13-16.

Berlyne, D. E. (1971). Aesthetics and psychobiology. New York: Appleton-Century-Crofts.

Christensen-Szalanski, J. J. (1984). Discount functions and the measurement of patients' values: Women's decisions during childbirth, *Medical Decision Making, 4*, 47-58.

Csikszentmihalyi, M. (1996). *Creativity: Flow and the psychology of discovery and invention*. New York: HarperCollins.

Cupchik, G. C. (1992). From perception to production: A multilevel analysis of the aesthetic process. In G. C. Cupchik & J. Laszlo (Eds.), *Emerging visions of the aesthetic process* (pp. 83-99). Cambridge: Cambridge University Press.

Davelaar, E. J, Haarmann, H. J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short-from long-term recency effects: Testing a neurocomputational model of list memory. *Memory & Cognition, 24,* 323-334.

DiGirolamo, G. J., & Hintzman, D. L. (1997). First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin & Review*, *4*, 121-124.

Duemer, J. (1991). William Wordsworth. In J. R. Greenfield (Ed.), *Dictionary of literary biography, Volume 107: British romantic prose writers, 1789-1832, First series.* Farmington Hills, MI: The Gale Group.

Epstein, R. (1985). The nature of publications in academic vitae. *American Psychologist, 40*, 240-241.

Feist, G. J. (1999). The influence of personality on artistic and scientific creativity. In R. J. Sternberg (Ed.), *Handbook of human creativity* (pp. 273-296). New York: Cambridge University Press.

Gentile, C. A., & Kaufman, J. C. (2002a, January). *Creative writing: "Sacred cow" or scorable construct?* ETS Workshop Series. Educational Testing Service, Princeton, New Jersey.

Gentile, C. A., & Kaufman, J. C. (2002b, March). *Re-analysis of creative writing data from the 1998 NAEP Classroom Writing Study.* Invited presentation to the National Assessment of Educational Progress Special Interest Group at the American Education Research Association, New Orleans, Louisiana.

Gilbert, D. (2007). *Stumbling on happiness*. London, UK: HarperPerennial.

Hayes, S. C. (1983). When more is less: Quantity versus quality of publications in the evaluation of academic vitae. *American Psychologist, 38*, 1398-1400.

Hennessey, B. A., & Amabile, T. M. (1999). Consensual assessment. In M. Runco & S. Pritzker (Eds.), *Encyclopedia of creativity* (pp. 347-359). New York: Academic Press.

Holmberg, D., & Holmes, J. G. (1987). Reconstruction of relationship memories: A mental models approach. *Personality and Social Psychology Bulletin, 13*, 228-238.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, A. (1993). When more pain is preferred to less: Adding a better ending, *Psychological Science, 4*, 401-405.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (in press). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal.*

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. (in press). Captions, consistency, creativity, and the consensual assessment technique: New evidence of validity. *Thinking Skills and Creativity.*

Kaufman, S. B., & Kaufman, J. C. (2007). Ten years to expertise, ten more to greatness: An investigation of modern writers. *Journal of Creative Behavior, 41,* 114-124.

Kipling, R. 1899. *The white man's burden.* London: Printed for private circulation.

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95,* 489-508.

Miller, J. K., Westerman, D. L., & Lloyd, M. E. (2004). Are first impressions lasting impressions? An exploration of the generality of the primacy effect in memory for repetitions. *Memory & Cognition, 32,* 1305-1315.

Morewedge, C. K., Gilbert, D. T., & Wilson, T. D. (2005). The least likely of times: How memory for past events biases the prediction of future events. *Psychological Science, 16*, 626-630.

Murdock, B. B., Jr. (1967). Recent developments in short-term memory. *British Journal of Psychology*, *58*, 421-433.

Simonton, D. K. (1977). Creative productivity, age, and stress: A biographical time-series analysis of 10 classical composers. *Journal of Personality and Social Psychology, 35*, 791-804.

Simonton, D. K. (1985). Quality, quantity, and age: The careers of ten distinguished psychologists. *International Journal of Aging & Human Development, 21*, 241-254.

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*, 66-89.

Simonton, D. K. (2000). Creative development as acquired expertise: Theoretical issues and an empirical test. *Developmental Review*, *20*, 283-318.

Simonton, D. K. (2003). Expertise, competence, and creative ability. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 213-240). Cambridge, UK: Cambridge University Press.

Smith, L. F., & Smith, J. K. (2006). The nature and growth of aesthetic fluency. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 47-58). Amityville, NY: Baywood.

Direct reprint requests to:

Scott Barry Kaufman
Department of Psychology
Yale University
Box 208205
New Haven, CT 06520-8205
e-mail: scott.kaufman@yale.edu