Discernment and Creativity : How Well Can People Identify Their Most Creative Ideas?*

By: Paul J. Silvia

Silvia, P. (2008, August). Discernment and creativity: How well can people identify their most creative ideas?. *Psychology of Aesthetics, Creativity, and the Arts*, 2(3), 139-146. DOI:10.1037/1931-3896.2.3.139

Some ideas should never see the light of day. It shouldn't surprise us that someone thought of selling artificial testicles for neutered dogs, measuring the emotions of vegetables, or drinking urine to treat cancer: we all have some misses along with our hits. What is more startling is that the creator thought the idea was a hit, that it was good enough to refine, develop, and present to the world at large. Discernment—the ability to evaluate the creativity of one's ideas—is an important part of theories of creativity. Sociocultural theories distinguish between having an idea, which is easy, and developing an idea so that the domain's gatekeepers and audience accept it, which is hard (Sawyer, 2006; Sternberg, 2006). Cognitive theories contrast creating ideas (divergent thinking) with evaluating and revising ideas (convergent and evaluative thinking; Cropley, 2006; Runco & Smith, 1992). Darwinian theories distinguish between processes that generate a lot of ideas and processes that selectively retain the best ideas (Simonton, 1999).

But compared with the massive literature on idea generation, divergent thinking, and creative production, discernment has received relatively little attention from researchers (Kozbelt, 2007). The psychology of creativity knows much more about how people generate ideas than about how they judge their own ideas. The present research thus explores how well people can judge their creative ideas and if there are reliable individual differences in discernment. [1] After a critical look at the idea of accuracy in creativity judgments, I develop a method for (1) estimating the overall level of discernment in a sample and (2) predicting who is more discerning than others. A multilevel latent-variable analysis of responses to divergent thinking tasks examined how people's choices of their best responses covary with the judgments made by raters.

Can Creativity Judgments Be Accurate or Inaccurate?

Before reviewing past research, we must consider whether people's judgments of their ideas can be considered accurate or inaccurate. Research on discernment has made claims about accuracy—such as the concept of *evaluative accuracy* (e.g., Runco & Dow, 2004; Runco &

∗ Acknowledgement: These data are from the Creativity and Cognition project, which was first described in Silvia et al. (2008). Chris Barona, Josh Cram, Karl Hess, Jenna Martinez, and Crystal Richard deserve thanks for collecting and scoring the data. Researchers interested in reanalyzing the data can download the data file and Mplus input files from the author's Web page.

Smith, 1992)—so this idea deserves a close analysis. Accuracy has been a pest in many areas of psychology, such as the accuracy of beliefs about the self (Silvia & Gendolla, 2001), the accuracy of impressions of other people (Funder, 1995), and the accuracy of judgments of one's future feelings (Gilbert & Ebert, 2002), but creativity research has not yet recognized accuracy's peskiness.

The problem hinges on how to define and quantify accuracy (Cronbach, 1955; Hastie & Rasinski, 1988). In an insightful analysis, Hastie and Rasinski (1988) evaluated several methodological approaches to assessing accuracy. Based on their review, they recommended an approach to accuracy that "involves a direct comparison between a judgment response and criterion value measured independently of the subject's judgments" (p. 196). An example is a person's self-reported judgment of his or her heart rate and a physiological assessment of the person's heart rate (e.g., Weisz, Balázs, & Ádám, 1988). The physiological value serves as a gold-standard criterion for the self-report. Research on accuracy goes awry when researchers choose an inapt criterion. In some cases, no criterion for accuracy exists, so judgments can be neither accurate nor inaccurate (Silvia & Gendolla, 2001).

For studying accuracy, Hastie and Rasinski (1988) recommended within-person designs to compare the judgment and criterion:

> If the judgments and criterion are measured quantitatively, then a judgment—criterion discrepancy index is frequently the basis for a summary of judgment accuracy. If the criterion and judgment are expressed in a categorical form (e.g., yes–no, present–absent), then usually several trials are performed and accuracy is indexed by the proportions of various types of correct and incorrect response. (p. 196)

Within-person designs allow estimates of each person's level of accuracy. These within-person effects can be aggregated for an estimate of the sample's overall accuracy. As a step further, however, researchers can model and explain variability in accuracy. With multilevel modeling (Hox, 2002), for example, researchers can use between-person factors (e.g., personality, intelligence, expertise) to explain why some people have higher accuracy levels than others. Thus, within-person designs can estimate a sample's overall level of accuracy as well as individual differences in accuracy.

Past Research on Discernment

How have creativity researchers measured accuracy? This section selectively reviews some representative research on the accuracy of creative judgments. Afterward, we'll consider a new methodological strategy for studying discernment.

Accuracy as Percentages of Hits

One method of assessing accuracy, developed by Runco and Smith (1992), uses simple hit rates (Wagner, 1993). In Runco and Smith's study, people completed three divergent thinking tasks. After the tasks, they were told to rate each response on a 1–7 scale, in which higher scores represented more creative responses. The people's ratings of their responses were then compared with whether their responses were original (unique within the sample) or popular (given by many people). Ratings of 6 and 7 were defined as original, and responses of 1 and 2 were defined as popular. With this binary coding, accuracy for original responses could be expressed as the hit rate. Accuracy in originality judgments was 37%, which means that 37% of the unique responses

were rated as 6 or 7. Intriguingly, originality scores correlated with accuracy scores—people with more original responses had more hits, suggesting that creative people tend to be discerning.

A later study (Runco & Dow, 2004) used a similar procedure. After completing divergent thinking tasks, people gave each response a letter grade (*A* through *F*). The responses were classified as unique or uncommon based on their frequency in the sample. Evaluative accuracy was quantified with hit rates for unique responses (i.e., the number of unique responses given a high score) and for uncommon responses (i.e., the number of uncommon responses given a high score). As before, divergent thinking scores correlated with accuracy scores: people who generated a lot of unique ideas had higher hit rates.

There are some subtle but serious problems with this method. First, simple hit rates are biased as markers of accuracy. Although not well known in creativity research, the bias of hit rates is well known in areas of research that study classification and recognition (see Wagner, 1993). Consider, for example, someone who gives all 10 of her responses an *A*, but only two responses were unique. The person's accuracy for original responses is 100%—each original response received an *A*—but the person was obviously indiscriminate and inaccurate. Simple hit rates don't account for biases in the tendency to use certain response categories; alternate metrics, such as Wagner's (1993) unbiased hit rate H_U, are necessary.

Second, the intriguing correlation between originality and accuracy—people with more creative ideas also judge them more accurately—is probably an artifact. It's widely known that uniqueness scores are highly correlated with *fluency*, the number of responses (Hocevar, 1979; Silvia, 2008; Silvia et al., 2008). And hits are correlated with fluency, too: the more responses, the more possibilities for a hit. Fluency thus confounds both originality and hits, so a correlation between originality and hits may be spurious. The pattern of results found by Runco and Dow (2004, p. 12) strongly suggests a confound. People completed three divergent thinking tasks; unusualness scores (number of responses given by only a few people) and accuracy scores (number of hits) were computed for each. The three accuracy scores were essentially uncorrelated with each other (average $r = .09$), and the three unusualness scores were modestly correlated with each other (average $r = .14$). But, within each task, the unusualness scores and accuracy scores correlated substantially (average $r = .37$). If the accuracy scores don't correlate with each other, then there's no shared construct common to them. Without a common construct, there's no substantive meaning to correlations between the accuracy scores and the uniqueness scores. This pattern suggests that the psychometric confounding of fluency, hits, and originality causes the correlation. [2]

And third, this procedure provides a good example of the criterion problem raised by Hastie and Rasinski (1988). In both studies, the judgment and criterion diverged. People were asked to rate their responses on a 1–7 scale or an *A*–*F* scale, not to judge whether a response would be unique or common within the sample. The mapping of ratings to criterion—6 or 7 equals original—is sensible, but the judgment and criterion nevertheless differ. Furthermore, the criterion for originality (infrequency in the sample) will shift with the sample size: more responses are original in a small sample than in a large sample (see Silvia et al., 2008). For participants to judge their responses accurately, they would need to know how many other people provided responses.

Accuracy as Discrepancy Between Self and Other Ratings

Another way to assess accuracy is to compute the average discrepancy between the participants' ratings and judges' ratings. A recent example of this approach is a study by Grohman, Wodniecka, and Kłusak (2006). The authors were appropriately circumspect regarding claims about accuracy; I describe their study because it's an interesting instance of one of psychology's oldest (and trickiest) methods of appraising accuracy (Cronbach, 1955). After completing three divergent thinking tasks, people rated the creativity of each response on a 1–7 scale and estimated the percentage of other people who gave the responses. Several judges then rated the creativity of each response on a 1–7 scale, and the percentages of people who gave each response were computed. Accuracy was estimated by (1) the differences between self ratings and the judges' ratings and (2) the differences between estimated percentages and observed percentages. For each index, difference scores near zero represent agreement between self ratings and criterion scores.

Much of Cronbach's (1955) classic critique of accuracy in person perception was devoted to this design. A central problem with using such a design to examine accuracy is the seductive but specious criterion: judges' ratings aren't really a gold standard. Judges are a source of variability, so they are more appropriately understood as a facet in the research design (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). First, judges' ratings are an inappropriate gold standard because judges and participants don't use the scale in the same way. Judges see all of the responses, but participants see only their responses, so judges and participants will have different representations of the central tendency and variability in responses. Second, it's hard to interpret the meaning of a self–other difference score. Consider a case in which participants give themselves higher ratings than the judges give. This disparity can mean, among other things, that (1) the judges are too strict; (2) the participants are too generous; (3) both the judges and participants are too strict, but the judges are stricter; (4) both the judges and participants are too lenient, but the judges are less lenient.

Third, judges can agree more strongly for some participants or for some items; fans of generalizability theory (Cronbach et al., 1972) will recognize these patterns as participant-by-judge and item-by-judge interactions. For example, judges might agree with each other more for terrible responses, thus rendering the criterion more or less valid across the range of scores and participants. And fourth, subtle regression-to-the-mean artifacts occur when participants give themselves high and low scores. Consider a person who had three bad responses and thus appropriately gave his responses 1s on a 1–7 scale. The raters can vary in only one direction—toward the mean—so any variation among the raters will make the participant seem too critical. Conversely, if a person gives her good responses 7s, the raters can vary in only one direction—toward the mean—so any variation among the raters will make the participant seem too generous.

Rethinking Accuracy

Thus far, attempts to quantify the accuracy of creativity judgments have been unsuccessful. The problem is not with the sophistication of past research but with the notion of creative accuracy itself. I suspect that most creativity researchers, in their heart of hearts (or brain of brains), would agree that there is no gold standard for creativity. Creative products probably do not have a true, innate level of creativeness—their creative worth is ultimately determined by complex

sociocultural and historical processes (Sawyer, 2006; Simonton, 1998). Without a gold-standard criterion, it is impossible to assess whether someone's judgments match the criterion scores. If creative judgments can be neither accurate nor inaccurate, what should researchers study? Instead of assessing accuracy, researchers can assess *extent of agreement*. In this approach, researchers assess the covariance of a person's judgments with criterion scores. The appropriateness of criterion scores is specified by relevant theory: they can be judges' ratings (Grohman et al., 2006), archival data (Kozbelt, 2007), or infrequency scores (Runco & Smith, 1992). It's worth pointing out that agreement is not a diminished, weak form of accuracy. Accuracy is a special case of agreement, the case in which the criterion scores come from a gold standard.

A good example of this methodological strategy is Kozbelt's (2007) analysis of Beethoven's discernment. Kozbelt collected all of Beethoven's self-critical statements; each statement was classified as positive or negative. Beethoven's judgments were then compared with markers of the contemporary critical consensus on the compositions (e.g., performance popularity, critics' ratings). Discernment was estimated as the covariance between self-ratings and archival criteria. Overall, Beethoven's statements about his work strongly predicted their critical success, indicating that Beethoven could discern his good work from his weaker work. And intriguingly, the extent of agreement increased over Beethoven's career, which suggests that he became more discerning with experience.

The Present Research

The present research has two goals. The first goal is to explore creative discernment. Overall, how well can people pick their best ideas? Are some people more discerning than others? Are creative people more discerning? People completed four divergent thinking tasks, for which they were instructed to try to be creative. After each task, they picked their "top two" responses, the two that they thought were the most creative. Three judges then independently rated all of the responses, using a 1–5 scale. Did people's judgments of their best ideas agree with the judges' ratings of their ideas? Was the level of agreement higher for some people than for others? Did creative people have higher agreement? The second goal is to demonstrate a fruitful methodological and analytical strategy for studying discernment. I used multilevel latent-variable models to estimate the typical level of discernment (i.e., the within-person relationship between people's top-two decisions and the judges' ratings) and to explain variance in within-person relationships using between-person predictors.

Method

Participants

The data for this study comes from the Creativity and Cognition project, which was first described in Silvia et al. (2008). The sample consists of 226 people (48 men, 178 women) enrolled in General Psychology at the University of North Carolina at Greensboro.

Procedure

People completed a wide range of measures of personality, cognition, attitudes, training, and demographics. The present analyses are based on responses to four of Wallach and Kogan's (1965) divergent thinking tasks: two unusual uses tasks (uses for a brick and for a knife) and two instances tasks (instances of things that are round and things that make a noise). The

experimenter explained that the study was about creative thinking styles, that the tasks assessed creative ways of thinking, and that the participants should try to be creative and to come up with creative responses to the task. People had 3 minutes for each task.

After each task, the experimenter asked people to read their responses and then circle the two responses that they thought were their most creative ones. People could take as much time as they wished for this decision, but in practice people quickly chose their top two responses. Responses chosen as a top-two response received a score of 2; all responses not chosen received a score of 1.

Divergent thinking tasks are controversial in the psychology of creativity (see Sawyer, 2006; Weisberg, 2006). In this study, the divergent thinking tasks are simply a context for people to generate and evaluate ideas. We asked people to try to be creative and then to pick their two most creative ideas. Other approaches, such as asking people to write captions for photographs (e.g., Kaufman, Lee, Baer, & Lee, 2007) and then to pick the best captions, could use the same statistical approach and ought to work equally well.

Each response was rated independently by three undergraduate research assistants; the instructions are available in Appendix 1 of Silvia et al. (2008). Each rater gave each response a score from 1 (*not at all creative*) to 5 (*highly creative*). The raters balanced a response's originality, remoteness, and cleverness, following Guilford's classic analysis of originality (Wilson, Guilford, & Christensen, 1953). In general, responses received higher scores when they were uncommon, when they were far from the typical use or instance, and when they were interesting, although strengths in one dimension could compensate for weaknesses in the others. The responses were sorted alphabetically within each task, so the raters were unaware of who gave the response, the person's other responses for that task, the response's serial order within the set, the person's fluency score for the task, and whether the response was picked as a top-two response. The raters read all of the responses before rating them, and they rated the responses independently of each other.

Complex issues surround measuring creativity with subjective ratings. Interested readers can consult comments on the target article that developed this subjective method—provided by Baer (2008); Kim (2008); Kogan (2008); Lee (2008); Mumford, Vessey, and Barrett (2008); and Runco (2008)—along with the rejoinder (Silvia, Winterstein, & Willse, 2008).

Measures of Between-Person Individual Differences

To measure between-person individual differences, we examined the Big Five domains of personality. Of the five domains, *openness to experience* was the most interesting. Many studies have found that openness is an important variable in art, aesthetics, and creativity (McCrae, 1987, 2007). For example, openness predicts better divergent thinking and more creative accomplishments (e.g., Carson, Peterson, & Higgins, 2005; Feist, 1998, 2006; King, Walker, & Broyles, 1996; McCrae, 1987; Silvia et al., 2008). We measured openness and the other four domains (extraversion, neuroticism, agreeableness, and conscientiousness) with three scales: the 60-item NEO Five Factor Inventory (Costa & McCrae, 1992), a 50-item scale from the International Personality Item Pool (IPIP) scale (Goldberg et al., 2006), and the 10-item Brief Big Five (BBF) scale (Gosling, Rentfrow, & Swann, 2003). Each scale used a 1–5 format.

Results

Model Specification

Figure 1 depicts the model that was estimated. The data have a two-level structure: a within-person level (the level of top-two choices and ratings) is nested in a between-person level (the level of individual differences). Multilevel modeling can simultaneously estimate within-person and between-person effects, thereby accounting for the interdependence of the observations (Hox, 2002; Silvia, 2007). Although typical multilevel models have only observed variables, this study estimated a general two-level model with both latent and observed variables (Muthén, 1997, 2002). The creativity ratings were modeled as a latent Creativity Ratings variable indicated by the scores of the three raters; the variance was set to 1. For the Big Five variables, the three scales served as indicators; for each, the path to the NEO scores was set to 1. Top-two choices, the within-person variable, was modeled as categorical; 23.3% of responses were chosen as top-two responses. All other variables were continuous.

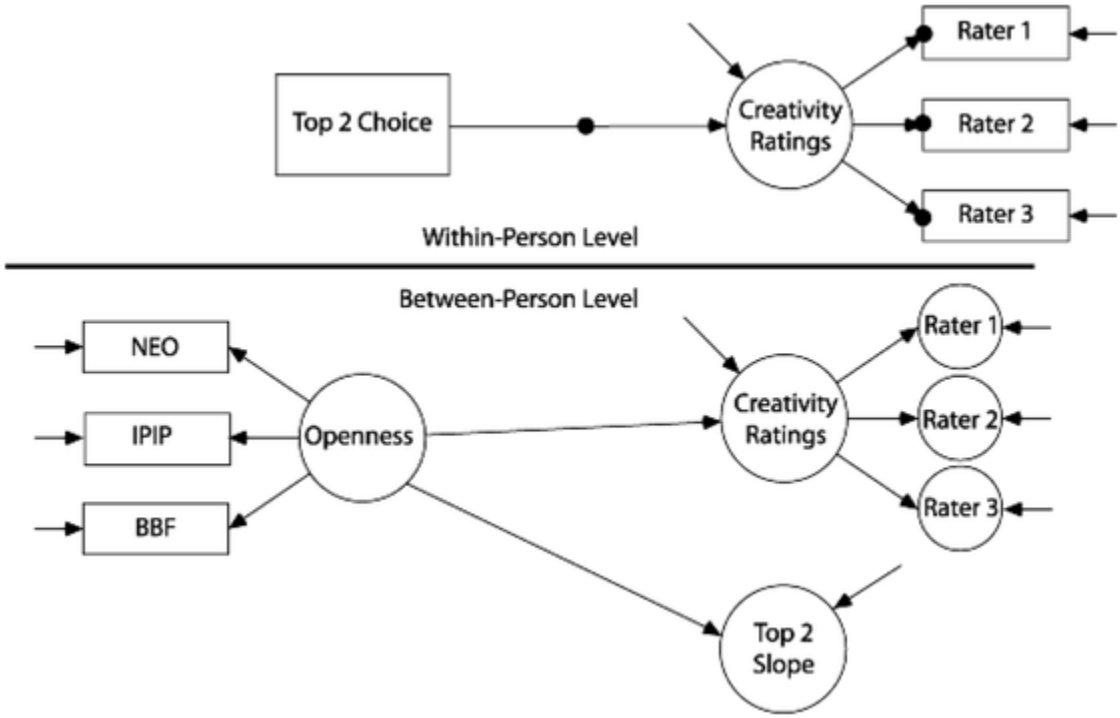

Figure 1. Depiction of the multilevel latent variable model.

At the within-person level, people's top-two choices predicted the latent Creativity Ratings (see Figure 1). The filled circle on the path indicates that the top-two slope was modeled as a random effect; the value of the slope was thus allowed to vary between people. Likewise, the filled circles for the scores of Raters 1, 2, and 3 indicate that these were modeled as random intercepts. At the between-person level, the random slope and intercepts appear as latent variables with values that vary across participants. Variation in the intercepts and in the top-two slopes was modeled as a function of between-person predictors. The analyses were conducted with Mplus 5, using maximum likelihood estimation with robust standard errors. All of the coefficients are unstandardized.

The within-person measurement model for Creativity Ratings showed that agreement between the raters was good. As expected, the paths for the first (b = .399, SE = .019, z = 21.2), second (b = .130, SE = .012, z = 11.3), and third (b = .514, SE = .021, z = 24.1) rater were positive and significant. Stated differently, the raters' scores were caused by the latent creative quality of the responses.

Overall Discernment: Did Top-Two Choices Predict Ratings?

How discerning was the sample as a whole? To assess the agreement between people's top-two choices and the judges' ratings of their responses, we can estimate the overall within-person relationship between participants' top-two judgments and the latent Creativity Ratings variable. This effect was significant, b = .77, SE = .046, z = 17.0, $p < .0001$, which indicates that the sample as a whole was discerning: people's choices of their best two responses covaried strongly with judges' ratings of the responses. [3]

Individual Differences: What Predicted Variance in Agreement?

Were some people in the sample more discerning than others? Were people high in openness to experience more discerning? To examine these questions, we next consider the effects of the between-person predictors on the top-two slopes. The paths to the top-two slopes indicate the degree to which the predictors explain variability in the slopes. Three models were estimated. In the first model, only openness to experience was included as a between-person predictor; Figure 1 depicts this model. In the second model, all five of the Big Five domains were included as predictors. In the third model, fluency was included as a predictor. Table 1 summarizes the findings of the first two models.

Table 1
Summary of the Multilevel Relationships

		Top-two slope			Creativity ratings		
Model	Predictor	*b*	*SE*	*p*	*b*	*SE*	*p*
1	Openness	.349	.139	.012	.926	.311	.003
2	Openness	.436	.188	.020	1.239	.449	.006
	Extraversion	−.095	.142	.504	−.282	.302	.351
	Neuroticism	.155	.098	.111	−.349	.193	.070
	Agreeableness	.302	.111	.006	−.067	.233	.772
	Conscientiousness	−.148	.087	.089	−.521	.173	.003

Note. b = unstandardized regression coefficients. Dividing b by its SE provides a critical ratio that approximates the z distribution in large samples. The p values are two-tailed.

Summary of the Multilevel Relationships

Openness to experience

Openness to experience significantly predicted creativity ratings: people high in openness had higher scores overall, consistent with past research on openness and divergent thinking (e.g., McCrae, 1987). Openness also predicted top-two slopes: people high in openness had higher slopes, indicating that their top-two judgments covaried more strongly with the scores given by

the raters. Both effects were similar regardless of whether the other four personality domains were included in the model (see Table 1). Taken together, people high in openness to experience had better responses and were more discerning when picking their best two responses.

Other domains

The other four domains were included as control variables, but some of their effects are interesting in their own right (see Table 1). Conscientiousness, in particular, has predicted low creativity in some past research (e.g., King et al., 1996). In the model with all five factors, high conscientiousness predicted lower creativity ratings and smaller top-two slopes. People high in conscientiousness thus not only gave less creative responses, but they made top-two judgments that agreed less strongly with the scores given by the raters. In some respects, then, conscientiousness and openness had opposing relationships with creativity and discernment: high openness, but low conscientiousness, predicted greater creativity and discernment. Additionally, agreeableness—a variable that has not received much attention in psychometric studies of creativity—significantly predicted top-two slopes.

Fluency

The final model added fluency. On average, people generated 8.18 uses for a brick, 8.12 uses for a knife, 10.82 instances of round things, and 13.49 instances of things that make a noise. Fluency was modeled as a latent variable with four indicators; the path to the brick scores was fixed to 1. Fluency's relationships with the Big Five variables were small—including it did not appreciably change the results in Table 1. High levels of fluency predicted lower creativity ratings (b = −.231, SE = .112, z = 2.07, p = .039): people who opted for quantity over quality had poorer responses, all else equal. (Similar effects appeared in Silvia et al.'s (2008) Experiments 1 and 2, with the reminder that these data overlap with data from Experiment 2.) What about discernment? High levels of fluency marginally predicted higher top-two slopes (b = .057, SE = .031, z = 1.82, p = .069), suggesting that people who generated a lot of responses were more discerning.

General Discussion

How well can people judge the creativity of their ideas? Many theories of creativity have proposed a role for the evaluation of ideas (see Sawyer, 2006; Sternberg, 2006). But research on evaluation lags far behind research on generation, despite some good efforts to understand how everyday people (Runco & Smith, 1992) and eminent creators (Kozbelt, 2007) judge their work. In the present research, we applied multilevel latent-variable models to explore how well people's choices of their best ideas agreed with judges' ratings of the quality of the ideas.

Generation and Evaluation as Distinct Creative Traits

Theories of creativity typically distinguish between idea generation and idea evaluation. Sternberg's (2006, p. 88) investment model, for example, proposes three intellectual skills important to creativity: a *practical–contextual* skill, the ability to "persuade others of…the value of one's ideas"; a *synthetic* skill, the ability "to see problems in new ways and to escape the bounds of conventional thinking"; and an *analytic* skill, the ability "to recognize which of one's ideas are worth pursuing and which are not." In Sternberg's model, these thinking skills are dimensions of individual differences: people can vary on all three skills.

In the present research, we found clear evidence for individual differences in discernment, which resembles the analytic thinking skill in the investment model. People on average were discerning when picking their best responses, but some people's choices more strongly agreed with the judges' ratings. People higher in discernment had traits that characterize creative people (King et al., 1996; McCrae, 1987; Silvia et al., 2008): they were higher in openness to experience. These findings indicate that generation skills and evaluation skills are distinct but correlated traits—people high in one tend to be high in the other. Creative people thus can come up with good ideas and discern which ideas are the good ones.

The notion of discernment as a trait of creativity expands the controversy over domain-general creative abilities (Kaufman & Baer, 2005). Most of the controversy has concerned domain-general ideational abilities, such as divergent thinking (Plucker, 2004, 2005). If judging ideas is a distinct trait (Sternberg, 2006), however, then we can ask if the ability to judge the creativity of one's ideas is general across domains or specific to domains in which one has creative skill and training. This question will likely be answered by longitudinal studies of the growth of creative expertise. Perhaps training provides strategies for selecting one's best ideas just as it fosters generating good ideas. Kozbelt's (2007) study of Beethoven, for example, found that Beethoven's self-judgments predicted critical success more strongly over time, suggesting that he developed in discernment.

Choice as a Marker of Self-Evaluation

This study used choice as the measure of self-judgment—people had to pick the two responses that they thought were the most creative. Past research has typically asked people to rate their responses (Grohman et al., 2006; Runco & Smith, 1992), not to choose a subset of them. Methodologically, both approaches have merit: ideally, researchers could measure both ratings and choices and then form a latent variable from them. Choice deserves attention in future research because discernment in real-world creativity typically entails making *yes–no* decisions. For example, a psychologist might have six ideas for a book, but she can write only one at a time. She may rate each idea as creative, but to enact the idea she must choose one and set aside the others. Likewise, a researcher invited to contribute a chapter to an edited book can develop only one chapter out of the various ideas that might come to him, even if he likes (or dislikes) the ideas equally. A photographer submitting work to a juried exhibition can submit only a handful of her photographs, regardless of how many she likes. Readers fresh from the academic job market will recall that applicants for academic jobs can include no more than three or four publications; most applicants must thus select their most creative work. Forced-choice measures of creative judgment could offer insight into how people make such real-world creative judgments.

The Versatility of Agreement

In the Introduction, I proposed that researchers should avoid the notion of accuracy in creativity judgments. Few creativity theories would want to commit to the idea that ideas and products have a true level of creativity. And if no gold standard exists, then self-judgments of creativity can be neither accurate nor inaccurate. As an alternative, researchers can focus on the more general question of *agreement*, the extent to which people's judgments of their creativity covary with criteria. Agreement is empirically tractable and methodologically versatile. Researchers needn't choose self-judgments and criterion judgments that have the same scale, such as a 1–7

rating scale (Grohman et al., 2006). Self-judgments can be different kinds of ratings or choices; criterion scores can be ratings by judges or recognized critics, archival scores (e.g., ticket sales, citation counts), or any theoretically defensible marker of creativity. Both the predictors and outcomes can be modeled as latent variables, a more powerful method for multivariate data than the methods typically used in creativity research. For example, indicators of the creativity of classical music—critics' ratings, number of performances, appearance in major anthologies—correlate highly with each other (Kozbelt, 2005, 2007). Instead of analyzing each criterion separately, researchers can estimate models with a single latent "creativity" outcome.

Conclusion

The prevalence of tired, rewarmed ideas could make a cynic conclude that people are bad at separating their creative ideas from their unoriginal ideas. This study explored how well people can pick their best ideas. People could choose their best ideas better than chance: their choices covaried significantly with judges' creativity ratings. But some people—such as people high in openness to experience—were more discerning. Creative people are thus doubly skilled: they are better at generating creative ideas and at discerning which ones are the best.

Footnotes

[1] This article focuses on the judgment of one's own ideas rather than the judgment of others' ideas (Lonergan, Scott, & Mumford, 2004). Intrapersonal and interpersonal judgments of creativity probably differ in important ways (Charles & Runco, 2001), although the methodological approach developed in this article could be applied to both kinds of judgments.

[2] Runco and Dow (2004) also created a "percent original" score in an effort to control for fluency, but this score does not solve the problems with their results. Leaving aside the problem of analyzing ratios of highly correlated variables, there are two issues. First, the accuracy score is derived from simple originality hit rates, so it inherently contains information about raw originality scores and hence fluency scores. Second, the three accuracy scores fail to correlate with each other, so their correlations with originality scores (be they raw originality scores or percent-original scores) have a questionable interpretation.

[3] Afficionados of generalized latent variable modeling will recognize important differences between conventional multilevel models (as enacted in programs such as HLM) and a latent variable representation of a multilevel model (e.g., Muthén, 1997, 2002). For example, the within-person path from the top-two scores to the ratings is, strictly speaking, the intercept of the between-person latent random effect, not a within-person path.

References

Baer, J. (2008). Commentary: Divergent thinking tests have problems, but this is not the solution. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 89–92.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, *17*, 37–50.

Charles, R. E., & Runco, M. A. (2001). Developmental trends in the evaluative and divergent thinking of children. *Creativity Research Journal*, *13*, 417–437.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity.”*Psychological Bulletin*, *52*, 177–193.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*, 391–404.
Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review*, *2*, 290–309.
Feist, G. J. (2006). *The psychology of science and the origins of the scientific mind*. New Haven, CT: Yale University Press.
Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.
Gilbert, D. T., & Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology*, *82*, 503–514.
Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality assessment. *Journal of Research in Personality*, *40*, 84–96.
Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528.
Grohman, M., Wodniecka, Z., & Kłusak, M. (2006). Divergent thinking and evaluation skills: Do they always go together?*Journal of Creative Behavior*, *40*, 125–145.
Hastie, R., & Rasinski, K. A. (1988). The concept of accuracy in social judgment. In D.Bar-Tal & A. W.Kruglanski (Eds.), *The social psychology of knowledge* (pp. 193–208). New York: Cambridge University Press.
Hocevar, D. (1979). Ideational fluency as a confounding factor in the measurement of originality. *Journal of Educational Psychology*, *71*, 191–196.
Hox, J. (2002). *Multilevel analysis*. Mahwah, NJ: Erlbaum.
Kaufman, J. C., & Baer, J. (Eds.) (2005). *Creativity across domains: Faces of the muse*. Mahwah, NJ: Erlbaum.
Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity*, *2*, 96–106.
Kim, K. H. (2008). Commentary: The Torrance Tests of Creative Thinking have already overcome many of the perceived weaknesses that Silvia et al.'s (2008) methods are intended to correct. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 97–99.
King, L. A., Walker, L. M., & Broyles, S. J. (1996). Creativity and the five-factor model. *Journal of Research in Personality*, *30*, 189–203.
Kogan, N. (2008). Commentary: Divergent-thinking research and the Zeitgeist. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 100–102.
Kozbelt, A. (2005). Factors affecting aesthetic success and improvement in creativity: A case study of the musical genres of Mozart. *Psychology of Music*, *33*, 235–255.
Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music*, *35*, 144–168.
Lee, S. (2008). Commentary: Reliability and validity of uniqueness scoring in creativity assessment. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 103–108.
Lonergan, D. C., Scott, G. M., & Mumford, M. D. (2004). Evaluative aspects of creative thought: Effects of appraisal and revision standards. *Creativity Research Journal*, *16*, 231–246.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*, 1258–1265.
McCrae, R. R. (2007). Aesthetic chills as a universal marker of openness to experience. *Motivation and Emotion, 31,* 5–11.
Mumford, M. D., Vessey, W. B., & Barrett, J. D. (2008). Commentary: Measuring divergent thinking: Is there really one solution to the problem?*Psychology of Aesthetics, Creativity, and the Arts*, *2*, 86–88.
Muthén, B. (1997). Latent variable modeling with longitudinal and multilevel data. In A.Raftery (Ed.), *Sociological methodology* (pp. 453–480). Boston: Blackwell.
Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117.
Plucker, J. A. (2004). Generalization of creativity across domains: Examination of the method effect hypothesis. *Journal of Creative Behavior, 38*, 1–12.
Plucker, J. A. (2005). The (relatively) generalist view of creativity. In J. C.Kaufman & J.Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 307–312). Mahwah, NJ: Erlbaum.
Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 93–96.
Runco, M. A., & Dow, G. T. (2004). Assessing the accuracy of judgments of originality on three divergent thinking tests. *Korean Journal of Thinking and Problem Solving*, *14*, 5–14.
Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences*, *13*, 295–302.
Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation*. New York: Oxford University Press.
Silvia, P. J. (2007). An introduction to multilevel modeling for research on the psychology of art and creativity. *Empirical Studies of the Arts*, *25*, 1–20.
Silvia, P. J. (2008). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). *Creativity Research Journal*, *20*, 34–39.
Silvia, P. J., & Gendolla, G. H. E. (2001). On introspection and self-perception: Does self-focused attention enable accurate self-knowledge?*Review of General Psychology*, *5*, 241–269.
Silvia, P. J., Winterstein, B. P., & Willse, J. T. (2008). Rejoinder: The madness to our method: Some thoughts on divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 109–114.
Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., et al. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85.
Simonton, D. K. (1998). Fickle fashion versus immortal fame: Transhistorical assessments of creative products in the opera house. *Journal of Personality and Social Psychology, 75*, 198–210.
Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.
Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18*, 87–98.
Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*, 3–28.
Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity–intelligence distinction*. New York: Holt, Rinehart, & Winston.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts*. Hoboken, NJ: Wiley.
Weisz, J., Balázs, L., & Ádám, G. (1988). The influence of self-focused attention on heartbeat perception. *Psychophysiology*, *25*, 193–199.
Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*, 362–370.
Submitted: January 14, 2008 *Revised:* March 25, 2008 *Accepted:* March 25, 2008