Rejoinder: The Madness to Our Method: Some Thoughts on Divergent Thinking

By: Paul J. Silvia, Beate P. Winterstein, and John T. Willse

Silvia, Paul J., Winterstein, Beate P., Willse, John T., Rejoinder: The madness to our method: Some thoughts on divergent thinking. Psychology of Aesthetics, Creativity, and the Arts, 1931-3896, 2008, Vol. 2, Issue 2. DOI: 10.1037/1931-3896.2.2.109

Abstract:
In this reply, the authors examine the madness to their method in light of the comments. Overall, the authors agree broadly with the comments; many of the issues will be settled only by future research. The authors disagree, though, that past research has proven past scoring methods—including the Torrance methods—to he satisfactory or satisfying. The authors conclude by offering their own criticisms of their method, of divergent thinking, and of the concept of domain-general creative abilities.

Article:
It is unoriginal to start a reply by thanking the scholars who wrote comments, but sometimes originality is overrated: We appreciate the care that went into these detailed and insightful comments. But we hope that this reply does not degenerate into a stereotype, in which we defend each of our claims to the death—indeed, many of our claims are probably already dead. Instead, we would like to use the comments as a starting point for exploring some new ideas and complex issues that could not be squeezed into our overly long target article (Silvia et al., 2008). The first sections of this reply consider issues raised and inspired by each comment. The last section considers our own criticisms of our scoring method, of divergent thinking, and of the notion of domain-general creative abilities. By the end, we hope that readers of this exchange have a stronger sense of unresolved issues and future directions for divergent thinking research.

What About the Torrance Tests?
Kyung Hee Kim (2008) provided a nice overview of the Torrance Tests, which differ from other classes of divergent-thinking tests in some important ways. This comment offers a chance to consider some issues, strengths, and weaknesses of the Torrance Tests. Before doing so, though, we tackle a few minor points of clarification. Kim broke our objections down to four points; we modestly take issue with a couple of them.

First, as an aside, we didn't (and wouldn't) argue that divergent-thinking tests are too highly correlated with general intelligence. Perhaps Simonton's (2003) quotation in our introduction—used to illustrate a famous researcher's skeptical position—gave this impression. Regardless, we agree with Kim (2008), and disagree with Simonton, on this point; Kim's (2005) fine meta-analysis has settled this classic issue.

Second, we do believe that divergent-thinking tests have changed little since the 1960s. This, however, is merely our opinion, not an evidence-based argument. (Some things do not change because they are already good: The Likert scale, for example, has not changed much since Rensis Likert and Gardner Murphy developed it in the 1930s.) Certainly, some creativity tests have been renormed, and some scoring systems have been created, refined, or jettisoned. But the core psychometric structure of these tests seems essentially the same: Researchers ask for responses and then quantify performance on the basis of some form of uniqueness. This is ultimately an opinion—we will leave it to readers to develop their own opinions.

Third, we're aware that the Torrance Tests instruct people to be creative. We see this as a strength of Torrance,s approach. And fourth, we certainly believe that fluency scores and uniqueness scores are highly correlated—so high, in fact, that they are confounded. We thus would not agree with Kim,s (2008) interpretation of our first study: "Silvia et al. could not avoid the high positive correlations between fluency scores and uniqueness scores in their study although they intended to solve their perceived problem." Our aim was not to create a better, refined uniqueness score; instead, the results of our Study 1 provide one more demonstration of the confounding of fluency and uniqueness. Our attempt to solve this problem (among others) was the testing of subjective scoring methods.

Do the Torrance Tests solve the problems that other divergent thinking tests face? The Torrance approach solves a few, we think. Instructing participants to be creative is essential to valid scores, in our view. The Torrance approach also uses a different sense of uniqueness: People receive points for "not unoriginal" responses, so the threshold for originality does not shift with sample size. Unlike the traditional Wallach and Kogan (1965) scoring system or the threshold scoring systems described by Runco (2008), the Torrance approach avoids the "large sample penalty" that we described in out target article.

But there is still a huge problem, one of the central problems that motivated our research. Originality scores correlate too highly with fluency scores, particularly in the Torrance Verbal tests. According to the latest Torrance (2008) norms, the median correlation between originality and fluency is r = .88. Kim (2008) skirted this issue, which is generally skirted in discussion of the Torrance tests. We understand that Torrance and others have argued that creative people can generate a lot of responses, but this feels like an admission of psychometric defeat: The unacceptably high overlap between originality and fluency forces this conclusion. Originality and fluency simply are not distinctly assessed by the Torrance approach—the scores are interchangeable.

We find the high correlation between fluency and originality/ uniqueness to be too high. At this level of covariance, one score is simply a coarsened measure of the other: Interpreting them as unique, distinct constructs strikes us as folly. Campbell and Fiske's (1959) multitrait–multimethod (MTMM) approach to construct validation supports our assertion: Discriminant validity is necessary for meaningful distinctions between constructs. But we leave it to the gentle reader to draw his or her own conclusion. Does any reader find this correlation acceptable? Does anyone see this correlation and conclude "What a relief! For a moment, I was afraid that fluency and originality measured the same thing." In any other area of research, would a correlation around r = .90 be evidence for discriminant validity? The Torrance Tests solve some of the problems that other divergent thinking tests face, but they do not solve this serious, fundamental problem of measurement.

What about Torrance’s figural tests? We did not have the space to discuss them in our target article, but there are some interesting issues surrounding them. Most generally, can the figural tests compensate for some of the problems with the verbal tests? As Kim (2008) pointed out, some of the figural tests control for the number of responses, thus separating quality from amount. And the figural tests can be scored for additional dimensions of performance, thereby enriching the verbal tests. But here’s the rub: The Torrance verbal tests and figural tests correlate modestly, at best—they are not indicators of the same higher-order construct (i.e., creativity, creative potential, or divergent thinking). In the Torrance data analyzed by Plucker (1999), a latent Figural factor correlated only r = .36 with a latent Verbal factor. Clapham (2004) found an identical correlation (r = .36), eerily enough, in a study of observed rather than latent Verbal and Figural scores.

The modest correlation between the verbal and figural tests is not necessarily a problem: In multifacteted intellectual assessments, researchers expect (and desire) some components to be weakly related (e.g., vocabulary knowledge and spatial reasoning). But researchers do not then treat the components as interchangeable, nor do they use the strengths of one to compensate for the other. When the verbal tests are criticized, the figural tests should not be raised in response, and vice versa. If the tests do not substantially measure the same underlying construct, then they are not interchangeable. In MTMM terms, the verbal and figural tasks are

"different trait, different method," not "same trait, different method." Because they vary in trait (we suspect) and in method, it is hard to compare them.

Wrapping up, we disagree that the Torrance tests solve the pressing problems faced by researchers interested in assessing divergent thinking. Nevertheless, we agree that new scoring methods should be compared with the Torrance scoring methods. The validity of a scoring method is ultimately an empirical issue. Perhaps a Torrance-certified rater is interested in reanalyzing the responses?

What About Objective Scoring Methods?

Mark Runco (2008), too, offered much food for thought, based on his decades of work in divergent thinking. Although a reader of our target article and of his comment may not believe it, we agree with Runco about most of his points and his global approach. We differ, though, in how we think of construct validity and in our interpretation of past research. Runco pointed to threshold scoring methods, such as using a threshold of 5% or 10% for uncommon responses. Threshold scoring methods go back at least half a century, and they have some virtues—but how well do they work? Many scoring systems have some desirable features; Runco himself, over the years, has developed and tested approaches that overcome some of the limitations to the systems favored by Torrance and by Wallach and Kogan (1965). But how well do they work?

In our view, there is not a lot of work that suggests strong evidence for validity. We do not mean that there is no evidence; to the contrary, there are many studies that show small-to-medium positive effects that are independent of fluency. If we want tasks that yield small, positive effects (based on small-sample studies), then we have found our tasks. If we want tasks that yield scores of this nature, then these tasks are valid for the purpose of obtaining small, consistent effects. (We would be curious to see what a meta-analysis has to say; our interpretation of the literature is a subjective rating, although we think that Baer [2008], would agree with us.) Either divergent thinking has only small and medium effects, or these assessment methods need refinement. We suspect that divergent thinking's true effects are larger but obscured by measurement error.

We also doubt that 5% and 10% threshold scoring systems solve the sample-size problem faced by uniqueness scoring. For a given participant, the chance of having a unique response (i.e., a response that passes the 5% or 10% cutoff) increases as his or her number of responses increases. Top Two scores, in contrast, have modest and occasionally negative correlations between creativity and fluency.

Moreover, there is little work that compares several scoring methods against each other; Mark Runco (2008) has done most of it, and this is the kind of research that needs to be done. We should point out that researchers can obtain our raw data, including the thousands of raw responses and rater scores. These data are ideal for comparing scoring systems. Researchers can rescore the responses using the Torrance approach, threshold approaches, and the percent-original system described by Runco. With these scores, we can compare the evidence for the validity of the scoring systems. Which system yields stronger relationships?

And although we sound repetitive, we again point out that the correlation between fluency and originality is extremely high, around r = .90. If large-sample studies—such as Torrance,s (2008) normative data and Wallach and Kogan's (1965) data (Silvia, 2008)—show such correlations, the lower correlations in small-sample studies are probably underestimated. No statistical control can be done here: There is too little unique variance. We thus agree and disagree with Runco,s (2008) view that "[a]s a matter of fact, the unique variance of originality and flexibility scores from divergent thinking tests are reliable, even if fluency scores arc statistically controlled." We agree, because our studies showed strong partial effects of creative quality (raters, scores); and we disagree, because few past studies have shown medium or large partial effect sizes. This is a matter of interpretation for now, but we would like to see a meta-analysis.

Raters and Their Discontents

The issue of raters and the validity of their scores received a lot of attention in the comments. The objectivity of scoring methods is important to many researchers (e.g., Runco’s [2008] comment). We should point out that

raters sneak through the back door of many ostensibly objective studies. Whenever researchers screen, drop, pool, compare, collate, or evaluate responses, they are engaging in subjective rating. They may be doing so with only one rater, however, so variance due to raters cannot be modeled. For example, it is common to drop bizarre responses—but bizarre according to whom? Similarly, is the response "make a brick path" the same as "make a brick sidewalk"? This decision is easy, but it isn,t objective in the sense favored by proponents of objective scoring. As we pointed out in our article, validity is what is important about scores, not objectivity, apparent or otherwise.

Beyond this point, we agree with the comments that raised questions about how raters are selected, trained, and evaluated. Without a doubt, these issues are complex. Raters will not always agree, and it seems likely that the disagreement will be based on systematic rater-level variables (e.g., expertise, intelligence, creativity, and discernment). What traits or experiences, if any, are necessary to rate effectively? What is the meaning of expertise in rating such responses? Based on the success of computer-based scoring in writing assessment, should we develop "expert system" programs to score the responses, thereby avoiding human raters altogether?

In his comment, John Baer (2008) offered a strong defense and a good analysis of the consensual assessment technique, the best-known method involving subjective ratings. We agree with his analysis, apart from one central quibble: Our research did not use the consensual assessment technique, obviously. Our target article mentions the consensual assessment technique in one paragraph—as an example of a method that uses subjective ratings—and our assessment method varies from it in key ways (e.g., using novice raters and choosing the best two scores). Our approach is thus not a use of consensual assessment gone horribly awry, but rather a new method that resembles consensual assessment in that it uses subjective ratings. We suspect that Baer is barking up the wrong but more thoroughly researched tree.

Nevertheless, the role of expertise in task scoring is unknown and clearly complex. Unlike tasks used in the consensual assessment technique, divergent-thinking tasks are not a domain of creative accomplishment—they are simple tasks intended to assess traits associated with individual differences in creativity. What, then, would a highly trained expert in divergent-thinking tasks look like? And if our novice raters were inappropriate choices of raters, then would the large effects be even bigger if we found "better" raters? Our studies worked well and found large effects, points that should not be overlooked when considering the soundness of the methods.

Soonmook Lee (2008), too, raised some insightful points about sampling raters and assessing their agreement. He pointed out that raters, like participants, are sampled: They can be more or less deviant, and the chance of a deviant rater increases as the number of raters increases. In this case, a couple of trained raters are more desirable than many independent raters; Kim (2008), too, raised this point in the context of the Torrance tests. Baer (2008), in his comment, would disagree: Independence of raters is central to the validity of consensual assessment, he argued. We agree that rater training is worth exploring. In our studies, we did not extensively train the raters: We wanted their scores to be largely independent, along the lines of the consensual assessment technique. As a result, our findings are probably close to the lower end of possible rater agreement. Extensive training aimed at enhancing agreement should be feasible—it,s a good idea for a research project.

In short, the validity of subjective ratings must be examined closely and considered seriously. Only the accumulation of research will provide evidence for or against the validity of raters; more likely, it will highlight the training or traits needed for consistent, valid ratings and thus provide guidelines for researchers. (Some of this research could involve rescoring the responses by using different raters and examining how well the groups of raters agreed.) We are intrigued, though, by the possibility of creating expert systems for scoring the responses. Creativity researchers are probably appalled by the notion that software can judge human creativity, but if nothing else, it is objective.

A Few Statistical Nuts and Bolts

Soonmook Lee (2008) offered some incisive thoughts about psychometric and statistical issues raised by our subjective scoring method. The first set of issues concerns the generalizability analyses. Our original draft had more information about the G and D studies, including a discussion of whether tasks ought to be treated as fixed versus random; most of this work was placed in a Web appendix. We agree that tasks are typically viewed as random and that the models for subjective scores and uniqueness scores differ in key ways. In the end, Study 1 struck us as too small to conclude that the tasks are fixed or random: Our tiny study could not sustain our conclusions. We had only three tasks, one per type, so the tasks and types are confounded. We agree with Lee, but we think that a richer data set is necessary to settle this issue. Along the lines of G coefficients, Lee pointed out that the uniqueness scores would have fared better if they had had a rater facet instead of a single rater. This may be true, but the model of uniqueness scoring, as emphasized in Runco's (2008) comment, is that uniqueness scoring is objective and thus free of rater variance. The question of whether uniqueness scoring would improve with more raters is an empirical one. We expect some differences between raters in their uniqueness decisions, but the magnitude of the variance due to raters is unknown. Lee's claim that uniqueness scoring would have greater reliability than the other methods can be neither supported nor refuted with the data at hand.

We are skeptical of Lee's (2008) MTMM model of Study 1,s data, in which he concluded that a single scoring factor describes the data. The two subjective scoring methods are based on the same data: The Top Two scores are a subset of the Average scores, so they are highly correlated. The study has only around 75 cases, too, so it didn't surprise us that Lee's model failed to converge to an appropriate solution. For what it's worth, there are many well-known problems with estimating MTMM models with structural equation modeling (SEM; Wothke, 1996). The findings from Study 2 show clear differences between the subjective methods, and they provide evidence for their validity.

Regarding our second study, we share some of Lee's (2008) concerns about the higher-order Huge Two model. Although not raised by Lee, the model fit of the Big Five and Huge Two models is not great. Across two large samples, we have found neither the fine model fit nor the null Plasticity—Stability correlation found by other researchers (e.g., DeYoung, 2006)—perhaps it is a Southern thing. Another issue, though, is that the lower-order Stability factors relate differently to divergent thinking: Agreeableness has a positive effect, but Conscientiousness has a negative effect. The positive effect of Stability confuses these conflicting lower-order effects. Lee noted that the effect of Stability on divergent thinking was not significant. This effect is significant with some methods of estimating standard errors but not with others. Because we lacked a strong reason to use, for example, maximum likelihood with first-order derivatives instead of common maximum likelihood, we used common maximum likelihood for all of our models. This case is a good example of the value of effect sizes, which are stable across model estimation in a way that p values are not.

Baer (2008) proposed that subjective scoring methods suffer from a lack of standardization: Researchers cannot obtain scores that could be compared across studies because of differences in samples and raters. Basic research typically does not aspire to precise point estimates of a person's trait level. Nevertheless, the fields of test-equating and item-response theory—both mature areas in psychometric theory—would disagree with Baer. If different samples and raters complicate comparability, then what should we think about adaptive tests (e.g., the Graduate Record Examination) in which people receive the same score despite completing different items?

More generally, Mumford, Vessey, and Barrett (2008) pointed out limitations in the scope of evidence for validity, based on Messick,s (1995) model of validity; Baer (2008), too, questioned the value of the evidence for validity that we reported. We agree, of course; only so much can be accomplished in one study or in one article, particularly when the study involves a large sample of people completing time-consuming tasks. The extensions and elaborations that they suggest strike us as good ideas. Perhaps this is another example of Cronbach,s (1957) two disciplines of scientific psychology—our psychometric approach and Mumford et al.,s (2008) experimental cognitive approach—talking past each other.

We think Mumford et al. (2008) and Baer (2008) are perhaps too dismissive. Within the model of validity advocated by Messick (1995), validity is not a binary feature of a study, an idea, or an assessment tool.

Researchers gain evidence in support of validity over time. Certainly, researchers can quibble with how much evidence there is so far, but validity is not a simple thing you can get from a single study. To date, we have strong evidence for reliability (one kind of evidence for validity) and evidence for associations with domains of personality and college majors, roughly classified. This evidence is part of the field,s broad, half-century interest in personality and creativity, so it is connected to a meaningful stream of theory and research. With regard to validity, however, our studies are an early word, not the last word.

Domain General Traits and Potentials

Perhaps the thorniest problem in divergent-thinking research is whether divergent-thinking tasks measure a domain-general trait, be it creative cognition, creative potential, or simply creativity. The tradition associated with Guilford (1967), Wallach and Kogan (1965), and Torrance (2008) implies a general trait of creativity. Modern research on divergent thinking agrees (e.g., Plucker, 2005), although the field of creativity is split over the issue of domain-general traits, as Baer (2008) noted in his comment (see also Kaufman & Baer, 2005).

In his comment, Nathan Kogan (2008) provided a historical perspective on the assessment of creativity. The Wallach and Kogan (1965) tasks remain hugely popular, but it appears that their theoretical backdrop has been lost. Kogan pointed out that the Wallach and Kogan approach was founded on an associative model of creativity; this model lent meaning to the scores. With regard to our approach, Kogan (2008) argued that without a theory of creative ideation, "the basis for individual differences in [divergent-thinking] responses in the Silvia et al. study remains obscure." When you're right, you're right. Our research does not delve into the causes of the differences between people, and this limits its contribution to a general model of creativity. (Here, we suspect that Mumford et al. [2008] would agree and that Runco [2008] and Kim [2008] would disagree.) Kogan,s comment highlights what has changed over four decades of research. Modern psychometric research, in our opinion, is less concerned with the why and how of divergent-thinking tasks. The tasks have been elevated to measures of a general trait related to creativity ability or potential. Wallach and Kogan's concern with process has been obscured by Torrance's (2008) concern with predictive validity and Guilford's (1967) concern with structure.

During the time that Wallach and Kogan (1965) were developing their research, the notion of domain-general creative abilities seemed sensible. In our research, we spoke of divergent-thinking tasks as measuring simply creativity. This draws the ire of many creativity researchers, but we may as well be candid. Nearly all research with divergent-thinking tasks presumes that these tasks measure a global trait of creativity. (This presumption, we think, is probably wrong—more on this later.) There is value in being straightforward, but most psychologists prefer to call spades "digging process actualizers" and "excavation implements." If people believe that they are measuring global creative ability, then they should call their construct creativity or creative ability.

The issue of reifying divergent-thinking tasks won't be solved by calling the tasks measures of creative potential, a phrase advocated by Runco (2008). If someone has a trait-like "potential"—a stable thing that varies across people, remains stable over time, and influences observable outcomes—then the potential appears to be a trait. Why not cut to the chase and call it creativity or creative ability? What is creative potential if not the tendency to behave creatively, such as by being creative in everyday life, pursuing creative goals, and having creative accomplishments? As an analogy, consider fluid intelligence. We could call it the ability to solve novel problems, or we could call it the potential to solve novel problems. Our predictions and assessment are the same, so we do not gain much by calling it potential.

Some New Signs of Madness

To add to the critical cacophony, we add our own criticisms, thereby rending whatever coherence this reply may have had.

Is There an Abstract Trait of Creativity?

The issue of trait structure is one of the many ways in which the study of intelligence has surpassed the study of creativity. Intelligence researchers have developed and tested sophisticated models of how cognitive traits relate

to each other (e.g., Carroll, 1993). Creativity researchers, in contrast, have not come as far. Where are the modern two-level or three-level models of creativity? For example, where is the creativity version of fluid intelligence versus crystallized intelligence? Where are the models of the full structure of creativity, ranging from specific abilities to abstract abilities? Guilford's (1967) Structure of Intellect model, although excessively complex, was a good start. Torrance's (2008) tests have Verbal and Figural tasks, which ostensibly measure the same higher-order construct. Mumford et al. (2008) noted eight creative skills, which could be lower-order creative factors. Sternberg (2006) proposed three creative abilities associated with developing ideas, judging ideas, and persuading others about the merits of ideas. There is not much research, however, that develops these models as structural models of individual differences in creative abilities.

The field of intelligence is a good role model. If an abstract trait like g exists, then it ought to have lower-order traits (e.g., fluid intelligence) that themselves may have lower-order traits (e.g., spatial reasoning): This is one meaning of a trait's abstractness. The trait of g is not measured directly: It's inferred from the covariance of lower-order traits. There is thus no direct measure of g, strictly speaking. Creativity research, in contrast, measures creative potential "directly" with ideational tasks, usually divergent-thinking tasks. It is not known whether these tasks are higher order or lower order—they have no known spot in a structure of creative abilities. The models we mentioned earlier make predictions about these structures, hut such models haven't been tested in the way that models of intellectual structure have been tested. For this reason, we agree essentially with Mumford et al. (2008), who proposed that there must be more to creative ability than divergent thinking, but we think that there has been more theory than research about what the other skills might be.

In a sense, we are arguing that the field of creativity does not know whether divergent thinking is an abstract, domain-general skill. The assertions and dismissals are probably both premature. Divergent thinking may be abstract, like g, but it may be lower order, like spatial reasoning or reading comprehension. The gap between intelligence research and creativity research is striking. Both fields started at similar spots, but the psychometric study of intelligence is probably two generations ahead of the psychometric study of creativity.

Is Divergent Thinking Important to the Creative Accomplishments of Experts?

Viewed broadly, the argument over domain-general creative traits involves two positions: there are (comments by Kim [2008], Runco [2008], and Kogan [2008]) or are not (Baer's [2008] comment) abstract traits that predict creative accomplishments in different domains. (Hybrid positions propose that both sides are true; e.g., Plucker, 2005.) We think a third position deserves serious consideration: There are abstract creative traits, but these traits predict only the emerging creativity of novices. The creative accomplishments of experts, in contrast, are detached from abstract traits such as divergent thinking. The psychology of divergent thinking might be the study of novice, everyday creativity—it might have little to offer the study of expert creativity.

Research on expertise has found that many traits predict performance less strongly as expertise develops. Experts have domain-specific strategies and extensive, organized knowledge; their performance is based more on these acquired expertise structures than on global resources such as fluid intelligence (Ackerman, 2007; Ackerman & Beier, 2003; Ericsson & Ward, 2007). Consider, for example, the creative domain of writing. Expert writers are strategic in their writing: They spend more time planning than novices do, they have problem-finding strategies that help them choose and develop a topic, and they have more domain knowledge to apply to the process of composition and revision (Bereiter & Scardamalia, 1987). Experts' writing is top down and knowledge based, whereas novices, writing is bottom up and planless. Traits such as fluid intelligence, verbal fluency, and divergent thinking probably discriminate good novice writers from weak novice writers, but strategies and knowledge discriminate the experts from the novices.

There are reasons to suspect, then, that the predictive power of divergent thinking is true only for beginners, such as schoolchildren, college students, and people who have not delved deeply into creative domains—in short, for nearly all of the samples studied in creativity research. As people train in creative domains, they learn knowledge and strategies that enable them to produce creative work, such as a domain's tacit rules, tactics for

judging and elaborating ideas, and core skills necessary for high-level performance. We thus agree with Kogan,s (2008) criticism of our tasks and of divergent thinking, in which he pointed out

> the gulf between [divergent-thinking] performance under time-limiting conditions, and the thought processes presumed to underlie real-world creativity. The latter take place over lengthy stretches of time, and generally involve an incubation period in which initially unrelated associations or images come together to solve a creative problem. (p. 101).

The nature of expert creativity is probably of a different kind than novice creativity. Experts are not merely faster or better than novices; they approach, find, and solve problems in qualitatively different ways.

The argument over whether there is a domain-general trait of creativity, therefore, may be misleading. There could be a domain-general trait, but this trait is probably important only to novice creativity. The creative products made by experts are probably unrelated to broad traits such as divergent thinking. This is a "third way" in divergent thinking that makes neither the denouncers (e.g., Baer, 2008) nor the boosters (e.g., Kim, 2008; Runco, 2008) of divergent thinking happy. So far, an expertise approach to creative performance has not received much attention, apart from broad proposals (e.g., Ericsson, 1999) and broad dismissals (e.g., Simonton, 2003). Recent work, however, suggests that an expert-performance approach to creativity deserves a closer look (e.g., Kozbelt, 2005, 2007; Weisberg, 2006). It may be a way of unifying the psychology of everyday creativity with the psychology of landmark innovation.

Concluding Thoughts

We think that there are strengths to divergent thinking and to our approach to measuring it: The evidence for reliability and validity is strong, and we continue to believe that our methods solve some assessment problems that other methods can't solve or haven't solved. Only future research will reveal the merits and weaknesses of our approach. Researchers interested in our system may be interested in a clever Web-based program, developed by Pretz and Link (in press), that eliminates most of the tediousness of assessing divergent thinking. We also are happy to share our data with researchers interested in developing and comparing scoring methods.

And, of course, there are weaknesses to our approach and to divergent thinking. Our research was motivated by the belief that divergent thinking, a proud variable in the psychology of creativity, deserves another chance-but probably only one more chance. The assessment of divergent thinking has not changed much in recent decades, and there are doubts about whether this construct is important to understanding major creative accomplishments. Divergent thinking may be an abstract creative trait, but not enough is known about the multilevel structure of creative abilities. And divergent thinking may be important only to the creativity of novices. It is likely that people outgrow traits such as divergent thinking as they acquire expertise in their creative domain-but, as always, only future research will tell.

References

Ackerman, P. L. (2007). New developments in understanding skilled performance. *Current Directions in Psychological Science, 16,* 235-239. Ackerman, P. L., & Beier, M. E. (2003). Trait complexes, cognitive investment, and domain knowledge. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 1-30). New York: Cambridge University Press.

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.

Bcreiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Mahwah, NJ: Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multi method matrix. *Psychological Bulletin, 56,* 81-105.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Clapham, M. M. (2004). The convergent validity of the Torrance Tests of Creative Thinking and Creative Interest Inventories. *Educational and Psychological Measurement, 64,* 828-841.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology, 91,* 1138-1151.
Ericsson, K. A. (1999). Creative expertise as superior reproducible performance: Innovative and flexible aspects of expert performance. *Psychological Inquiry, 10,* 329-333.
Ericsson, K. A., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory: Toward a science of expert and exceptional performance. *Current Directions in Psychological Science, 16,* 346-350.
Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.
Kaufman, J. C., & Baer, J. (Eds.). (2005). *Creativity across domains: Faces of the muse.* Mahwah, NJ: Erlbaum.
Kim, K. H. (2005). Can only intelligent people be creative? *Journal of Secondary Gifted Education,* /6, 57-66.
Kim, K. H. (2008). Commentary: The Torrance Tests of Creative Thinking 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.
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.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons, responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.
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.
Plucker, J. A. (1999). Is the proof in the pudding? Reanalyses of Torrance,s (1958 to present) longitudinal data. *Creativity Research Journal, /2,* 103-114.
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.
Pretz, J. E., & Link, J. A. (in press). The creative task generator: A tool for the generation of customized, Web-based creativity tasks. *Behavior Research Methods.*
Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. *Psychology of Aesthetics, Creativity, and the Arts,* 2, 93-96.
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., 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. (2003). Expertise, competence, and creative ability: The perplexing complexities. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 213-239). New York: Cambridge University Press.
Steinberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18,* 87-98.
Torrance, E. P. (2008). *Torrance Tests of Creative Thinking: Norms-technical manual, verbal forms A and B.* Bensenville, IL: Scholastic Testing Service.
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.
Wothke, W. (1996). Models for multitrait-multimethod matrix analysis. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 7-56). Mahwah, NJ: Erlbaum.