Assessing Creativity with Divergent Thinking Tasks: Exploring the Reliability and Validity of New Subjective Scoring Methods

By: Paul J. Silvia, Beate P. Winterstein, John T. Willse, Christopher M. Barona, Joshua T. Cram, Karl I. Hess, Jenna L. Martinez, & Crystal A. Richard

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (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. http://dx.doi.org/10.1037/1931-3896.2.2.68

TABLES AND FIGUES CAN BE FOUND AT THE END OF THE ARTICLE.

*****Note: Figures may be missing from this format of the document**

Abstract:

Divergent thinking is central to the study of individual differences in creativity, but the traditional scoring systems (assigning points for infrequent responses and summing the points) face well-known problems. After critically reviewing past scoring methods, this article describes a new approach to assessing divergent thinking and appraises its reliability and validity. In our new Top 2 scoring method, participants complete a divergent thinking task and then circle the two responses that they think are their most creative responses. Raters then evaluate the responses on a 5-point scale. Regarding reliability, a generalizability analysis showed that subjective ratings of unusual-uses tasks and instances tasks yield dependable scores with only 2 or 3 raters. Regarding validity, a latent-variable study (n = 226) predicted divergent thinking from the Big Five factors and their higher-order traits (Plasticity and Stability). Over half of the variance in divergent thinking could be explained by dimensions of personality. The article presents instructions for measuring divergent thinking with the new method. Keywords: creativity, divergent thinking, generalizability theory, validity, reliability, assessment

Article:

The study of divergent thinking is one of the oldest and largest areas in the scientific study of creativity (Guilford, 1950; Weisberg, 2006). Within the psychometric study of creativity—the study of individual differences in creative ability and potential—divergent thinking is the most promising candidate for the foundation of creative ability (Plucker & Renzulli, 1999; Runco, 2007). For this reason, widely-used

creativity tests, such as the Torrance Tests of Creative Thinking (TTCT), are largely divergent thinking tests (Kim, 2006).

Nevertheless, modern writings on creativity reflect unease about the usefulness of divergent thinking tasks. In their reviews of creativity research, both Sawyer (2006) and Weisberg (2006) criticize divergent thinking research for failing to live up to its promise: after half a century of research, the evidence for global creative ability ought to be better (see Plucker, 2004, 2005; Baer & Kaufman, 2005). While reviewing the notion of creativity as an ability, Simonton (2003, p. 216) offers this blistering summary of creativity assessment:

> None of these suggested measures can be said to have passed all the psychometric hurdles required of established ability tests. For instance, scores on separate creativity tests often correlate too highly with general intelligence (that is, low divergent validity), correlate very weakly among each other (that is, low convergent validity), and correlate very weakly with objective indicators of overt creative behaviors (that is, low predictive validity).

We believe that researchers interested in divergent thinking ought to take these criticisms seriously. Although we don't think that the literature is as grim as Simonton's synopsis implies, divergent thinking research commonly finds weak internal consistency and rarely finds large effect sizes.

Informed by the large body of research and criticism (Sawyer, 2006; Weisberg, 2006), researchers ought to revisit the assessment and scoring of divergent thinking. There are many reasons for observing small effects—including genuinely small effect sizes—but low reliability seems like a good place to start. Methods of administering and scoring divergent thinking tasks have changed little since the 1960s (Torrance, 1967; Wallach & Kogan, 1965), despite some good refinements and alternatives since then (Harrington, 1975; Michael & Wright, 1989). It would be surprising, given the advances in psychometrics and assessment over the last 40 years, if the old ways were still the best ways.

In this article, we examine an alternative method of assessing and scoring divergent thinking tasks. Our method is simply a combination of past ideas that deserve a new look, such as the necessity of instructing people to be creative (Harrington, 1975) and the value of subjective ratings of creativity (Amabile, 1982;

Michael & Wright, 1989). The first part of this article reviews the assessment of divergent thinking and considers psychometric problems with these methods. We then appraise the reliability and validity of two new scoring methods: judges rate each response on a 5-point scale, and the ratings are averaged across all responses (*Average scoring*) or across only the two responses that people chose as their best responses (*Top 2 scoring*). In Study 1, we examined the reliability of these scoring systems by applying generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In Study 2, we examine the validity of the scoring systems with a latent variable analysis of personality and creativity. Finally, we review the implications of this work and provide take-home recommendations for researchers interested in using the new methods.

Assessing Divergent Thinking

Divergent thinking is assessed with divergent thinking tasks, in which people generate ideas in response to verbal or figural prompts (Kim, 2006; Michael & Wright, 1989; Wallach & Kogan, 1965). In a typical verbal task, people are asked to generate unusual uses for common objects (e.g., bricks, knives, newspapers), instances of common concepts (e.g., instances of things are round, strong, or loud), consequences of hypothetical events (e.g., what would happen if people went blind, shrank to 12 inches tall, or no longer needed to sleep), or similarities between common concepts (e.g., ways in which milk and meat are similar). Divergent thinking tasks are thus a kind of fluency task: they assess production ability in response to a constraint (Bousfield & Sedgewick, 1944). But unlike letter fluency tasks (e.g., list as many words that start with *M* as you can) and semantic fluency tasks (e.g., list as many cities as you can), divergent thinking tasks intend to capture the *creative quality* of the responses, not merely the number of responses.

Uniqueness Scoring

The most common way of scoring divergent thinking tasks is some form of *uniqueness scoring*. In their classic book, Wallach and Kogan (1965) criticized past efforts to assess and score creativity (e.g., Getzels & Jackson, 1962). As an alternative, they recommended pooling the sample's responses and assigning a 0 or 1 to each response. Any response given by only one person—a unique response—receives a 1; all other responses receive a 0. This scoring method has several virtues. First, it can be done by a single rater. Second, it's easier

than methods suggested by Guilford, such as weighting each response by its frequency (e.g., Wilson, Guilford, & Christensen, 1953). Finally, it has a straightforward interpretation—a creative response is a unique response.

The Wallach and Kogan uniqueness index is popular in modern creativity research, in part because of the popularity of the Wallach and Kogan tasks. Alternative scoring methods, however, share the same psychometric model. The Torrance Tests, for example, assign points for responses that fall outside a normative sample's pool of common responses (Torrance, 2008), and the points are then summed for an originality score. Other researchers assign 1 point for responses given by fewer than 5% of the sample and 0 points for all other responses (e.g., Milgram & Milgram, 1976), and these points are summed. Despite their surface differences, the Wallach and Kogan uniqueness score and the Torrance originality score share the same psychometric model: people receive points for statistically uncommon responses, and these points are summed.

Problems With Uniqueness Scoring

Uniqueness scoring, in our view, has three fundamental limitations. Two have been known for several decades; a third we raise for the first time.

1. Uniqueness Scoring Confounds Fluency and Creativity

Critics of divergent thinking research point out that uniqueness scores (the number of unique responses) are confounded with fluency scores (the total number of responses). In Wallach and Kogan's (1965) original study, for example, the confounding was severe: a recent reanalysis found a relationship of ß = .89 between latent uniqueness and fluency factors (Silvia, in press). Torrance's method of assigning points for not-common responses has the same problem. The latest TTCT verbal manual (Torrance, 2008) reports a median correlation of $r = .88$ between originality scores and fluency scores. This confounding is inevitable because the likelihood of generating a unique response increases as the number of responses increases. The confounding of uniqueness and fluency is a problem for two obvious reasons. First, the quality of responses and the quantity of responses ought to be distinct, according to theories of creativity, so creativity assessment ought to yield distinct estimates of quality and quantity. Second, the level of confounding can be

so severe that researchers cannot be certain that uniqueness scores explain any variance beyond mere fluency scores.

Since the 1970s, researchers have discussed the fluency confound as a problem and have considered ways of handling it (Clark & Mirels, 1970; Dixon, 1979; Hocevar, 1979a, 1979b; Hocevar & Michael, 1979). Many variations of uniqueness scoring have been proposed, such as weighting each response by its frequency (Runco, Okuda, & Thurston, 1987), scoring only the first three responses (Clark & Mirels, 1970), or quantifying fluency as the number of non-unique responses (Moran, Milgram, Sawyers, & Fu, 1983). Although worthwhile attempts, these scoring methods have not always performed well psychometrically (Michael & Wright, 1989; Speedie, Asher, & Treffinger, 1971). Furthermore, variations on uniqueness scoring do not overcome the two other criticisms.

2. Statistical Rarity is Ambiguous

The interpretation of unique scores is not as clear as it initially seems. Creative responses are not merely unique: they must also be appropriate to the task at hand (Sawyer, 2006). Many unique responses are not creative, but they slip through the cracks of the objective scoring system. First, bizarre, glib, and inappropriate responses are hard to filter from the pool of responses. Any researcher who has implemented the 0/1 system knows that the line between "creative" and "random" is often fuzzy. Researchers will disagree, for example, over whether "a round cube" or "a roundhouse kick from Chuck Norris!" should be filtered as capricious,

inappropriate responses to a "things that are round" task. Second, mundane responses will slip through the cracks of the uniqueness scoring system, thereby reducing its reliability. For example, "make a brick path" is an obvious use for a brick, but it could be unique in the small samples typical of creativity research.

In short, the objective 0/1 system is not as objective as it seems: it will tend to give 1s to weird responses and to common responses that raters would judge as uncreative. Some evidence for this claim comes from research that compared the objective 0/1 coding with subjective 0/1 coding. Hocevar (1979b) had 4 raters score responses using a 0/1 unoriginal/original scale. The raters' uniqueness scores were substantially lower than the objective uniqueness scores, indicating that the raters had a higher criterion for judging uniqueness.

3. Uniqueness Scoring Penalizes Large Samples

One of the biggest problems with uniqueness scoring—and one not recognized to date—is that it imposes a penalty on researchers who collect large samples. For uniqueness scoring, the creativity of a response depends on the pool of responses: a response is more likely to be scored as unique in a small sample than in a large sample. The probability that a response will appear in the pool is a function of the number of people, so as sample size increases, the probability that two people will give the same response increases. For example, the response "door knob" as an instance of something round would be scored 0 in a large sample but could be unique in a small sample. As a result, the base rate of creativity goes down as the sample's size goes up. Stated differently, the criterion that a response must pass to be a unique response is too low in a small sample and too high in a large sample. Creative responses are thus harder to detect, in a signal-detection sense, in large samples.

An extreme example demonstrates our point. With a sample size of 1, all of the lone participant's responses are creative. With a sample size of 5, only the most grossly obvious responses receive 0s, and most responses will receive 1s. With a sample size of 100,000 people, however, a response must be highly creative (or merely bizarre) to receive a 1. As a result, most people will have 0s for all of their responses. Creativity is harder to detect in a sample of 100,000 than in a sample of 100, because the criterion for creativity is excessively high (1 in 100,000 versus 1 in 100). And uniqueness scores needn't reach an asymptotically low level: it's theoretically possible (although unlikely) in vast samples for no response to receive a 1. There's something perverse about a psychometric method that performs worse with large samples. Researchers shouldn't be penalized for collecting large samples, particularly researchers interested in the psychometric study of individual differences.[1]

Subjective Scoring of Creativity as an Alternative

What alternative scoring methods can overcome these three problems? We think that creativity researchers ought to reconsider the value of subjective scoring of divergent thinking responses. There's a long tradition of scoring creativity by having trained raters evaluate peoples' responses. In the earliest modern creativity research, Guilford's research team used raters to score some of their divergent thinking tasks. To assess the *cleverness* component of creativity, Guilford had people generate plot titles, which were then scored

on 1-5 scales by 2 raters (Christensen, Guilford, & Wilson, 1957) or on 0-6 scales by 3 raters (Wilson et al., 1953). To assess the *remoteness of association* component of creativity, people generated responses to a consequences task; the responses were scored on a 1-3 "remoteness" scale (Christensen et al., 1957). Since Guilford, many researchers have used subjective ratings of responses to divergent thinking tasks, such as scoring each response on a 1-5 scale (Harrington, 1975) or a 1-7 scale (Grohman, Wodniecka, & K#usak, 2006), scoring responses as high or low in quality (Harrington, Block, & Block, 1983), and scoring the full set of responses on a 1-7 scale (Mouchiroud & Lubart, 2001).

Subjective scoring of creativity—particularly Amabile's (1982) *consensual assessment technique*—has been popular for several decades in the study of creative products. The consensual assessment technique entails independent judges—ideally but not necessarily experts—rating products for creativity, based on the judges' tacit, personal meanings of creativity. Judges often show high consistency and agreement (Amabile, 1982; Baer, Kaufman, & Gentile, 2004; Kaufman, Gentile, & Baer, 2005; Kaufman, Lee, Baer, & Lee, in press). Expertise enhances agreement, but recruiting experts is probably more important for studies of real creative products than for studies of responses to divergent thinking tasks. The consensual assessment technique has worked in a wide range of contexts and samples, indicating that the subjective scores have sufficient validity (see Amabile, 1996).

Subjective ratings can overcome the three problems faced by uniqueness scoring. First, ratings should, in principle, be unconfounded with fluency: because the raters judge each response separately, generating a lot of responses won't necessarily increase one's creativity score. Second, bizarre, weird, and common responses that slip through the cracks of the uniqueness index ought to be caught by the subjective raters. A common use for a brick like "make a brick path," for example, will always get low scores from raters. Moreover, several raters can evaluate the creativity of bizarre and weird responses, which is an improvement over the 0/1 decisions made by a single coder. And third, subjective ratings ought to be independent of

sample size. Creativity is scored by the standards set by raters, not by the frequency of responses in a pool. The raters' standards ought to be the same regardless of the sample's size, so the base rates of subjectively-scored creativity shouldn't be artificially inflated or depressed for small and large samples.

In the present research, we developed a system for subjective scoring of creativity. Raters received definitions of creativity proposed by Guilford, which they used as a guide for judging the creativity of each response. We then evaluated two indexes of creativity derived from these subjective ratings. The first index, *Average scoring*, is a simple average of all of a person's responses to a task. If someone generated 9 uses for brick, for example, the person's creativity score is the average of the ratings of those 9 uses. The second index, inspired by a suggestion made by Michael and Wright (1989, p. 48), controls for the number of responses. After generating their responses, people circle the two responses that they feel are the most creative. The judges' ratings of the top two responses are averaged to form each person's creativity score for the task. This *Top 2 index* evaluates people's best efforts, in their own judgment, and it thus represents people's best level of performance when they are instructed to do their best.

Our studies examine both scoring methods, but we expected Top 2 scoring to perform better than Average scoring. First, by examining people's best efforts, the Top 2 approach is a form of *maximal assessment*: people are evaluated by the best level of performance they are able to achieve (Runco, 1986). Second, the Top 2 approach holds constant the number of responses on which people are evaluated, which is a nice psychometric feature. Some people will give more responses than others, but each person is judged on his or her best two responses. And third, in real-world creativity, picking one's best ideas is as important as generating a lot of ideas (Grohman et al., 2006; Kozbelt, 2007; Sawyer, 2006). The Top 2 index allows people to decide which of their responses are hits and which are misses.

Many psychologists are skeptical of subjective scoring, particularly when an ostensibly objective method is available. Several researchers have contended that subjective ratings are simply too idiosyncratic to be useful: raters disagree too much with each other, and each person has his or her own vague definition of creativity (see discussions by Michael & Wright, 1989, and Runco & Mraz, 1992). In our view, subjective scoring should be considered seriously. First, the idiosyncrasies of raters have been overstated: many studies show excellent

agreement in the subjective judgments of independent raters (Amabile, 1982; Baer et al., 2004; Kaufman et al., 2005). Second, agreement between raters can be enhanced by giving them clear instructions, by providing accepted definitions of creativity, and by training them in the scoring system. Finding low agreement isn't surprising when the raters aren't trained or instructed (e.g., Runco & Mraz, 1992). Third, variance associated with raters needn't be mere error—rater variance can be modeled, thus reducing overall error. And fourth, the merit of a subjective scoring system is an empirical question. What's important about scores is their reliability and validity, not their ostensible level of objectivity or directness (Webb, Campbell, Schwartz, & Sechrest, 1966). Whether subjective methods are better than objective methods is a matter for research, such as the present research.

The Present Research

The present research evaluated the reliability and validity of the two subjective scoring methods: Average scoring and Top 2 scoring. In Study 1, we conducted a generalizability analysis to estimate the variance in scores due to real differences between people and to differences between raters. Dependable scores would have most of the variance due to between-person differences in divergent thinking and much less variance due to the raters. For contrast, we compared the two subjective scoring methods with the Wallach and Kogan (1965) uniqueness index. In Study 2, we evaluated the validity of the scoring methods by conducting a large-sample latent-variable analysis of personality and divergent thinking. If the scores are valid, then we ought to be able to explain substantial variance in divergent thinking with theoretically important predictors of creativity, such as dimensions of personality (e.g., Openness to Experience) and lifestyle (e.g., choosing to pursue a college major related to the arts).

Study 1: The Dependability of Average Scores and Top 2 Scores

Generalizability theory (Cronbach et al., 1972; Shavelson & Webb, 1991) was chosen to examine the reliability of divergent thinking scores—or as generalizability theory (G-theory) puts it, the *dependability* of scores.[2] Unlike classical test theory (CTT), G-theory takes in account more than one type of measurement error within the same analysis—error is considered multifaceted. In CTT, for example, coefficient alpha estimates only how consistently items measure a construct. Generalizability analysis can estimate how consistently items

behave and raters behave, and it can take them both into account in the same coefficient. Generalizability analysis disentangles error by partitioning the variances that are accounted for by the object of measurement and by the defined facets of measurement error. Facets are major sources of variation, and the conditions under random facets are considered interchangeable. For example, if raters are a facet, then the researcher is willing to treat the raters as interchangeable. Facets besides rater and task, for example, could also be time limit, rating scale, and testing condition. But it is the researcher's task to determine, based on theoretical considerations and previous research findings, what types of measurement error are relevant for an instrument and its application. Equation 1 (Brennan, 2001) shows how variance components are decomposed in G- theory in a person-by-task-by-rater design. The observed score variance is partitioned into person variance, task variance, rater variance, person-by-task variance, person-by-rater variance, task-by-rater variance, and the confounded variance of person-by-task-by-rater variance and error.

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr). \quad (1)$$

By partitioning error associated with measuring divergent thinking, researchers receive guidance on how to improve the precision of the scores. Based on estimated variances associated with conditions of measurement (e.g., raters), the dependability can be increased by adding conditions to the facet that contributes most of the error. For example, if the analysis indicates that raters are a big source of inconsistency, then researchers can find out how many raters they need to get a desired dependability level or they could concentrate more efforts on rater training. This is conceptually similar to determining the increase in reliability in classical test theory by applying the Spearman-Brown formula for items. Generalizability analysis provides this information because G-theory can partition error variance attributable to separate sources. Based on this information, the analysis also offers estimates of dependability for modified measurement scenarios. For example, it estimates dependability for 2 raters and 4 tasks, 3 raters and 6 tasks, and any other possible combination. Researchers can then use these estimates when planning research.

Generalizability analysis can provide estimates of the dependability of instruments applied for norm-reference measurement as well as criterion-reference measurement. Specifically, if the goal is to compare

examinees on their divergent thinking scores against each other—in other words, if we are interested in their relative standing to each other—then the generalizability coefficient (G-coefficient), analogous to coefficient alpha, would inform us how dependable a score is. On the other hand, if we are interested in the absolute standing of an examinee to a criterion, or how much the observed divergent thinking score deviates from the true score, then we would want to know the phi-coefficient (phi coefficient) of a measure. The distinction between dependability for decisions on the relative standing of examinees and decisions on the absolute standing of examinees can be made because G-theory provides estimations for G-coefficients (for relative decisions) and phi coefficients (for absolute decisions). G-coefficients are higher than - coefficients because they consider different error terms in calculating the dependability. Because we are interested in the relative standing of examinees in relative decisions, the G-coefficient considers only error terms associated with interaction effects. On the other hand, when we are interested in absolute decisions, we must consider the main error effects as well. The greater error term for the denominator of the formula shrinks the - coefficient. The generalizability and - coefficients will be identical if there is no error associated with the main error effects.

Another differentiation that G-theory makes is between generalizability studies and decision studies. Whereas the generalizability study (G-study) provides the variance components and dependability coefficients associated with the study's measurement design, the decision study (D-study) estimates the variance components and dependability coefficients for alternative study designs. For example, our original design includes three raters and three tasks, and the G- study informs about the variance decomposition and the dependability coefficients. The D-study then provides estimates for alternative designs with different combinations of raters and tasks: 2 raters and 3 tasks, 4 raters and 3 tasks, 3 raters and 5 tasks, and so forth.

Method

Participants and Design

A total of 79 undergraduate students enrolled in General Psychology at the University of North Carolina at Greensboro (UNCG) participated as part of a research participation option. Two people were excluded because of substantial missing data, yielding a final sample of 77 (48 women, 29 men). The sample had a wide range of majors: the most common majors were fine arts and performing arts (12%), undeclared (9%), education (8%), and psychology (8%). *Divergent Thinking Tasks*

People arrived at the lab in groups of 3 to 8. After completing a consent form, they learned that the study was about the psychology of creativity. From the beginning of the study, the experimenter emphasized that the researchers were interested in how people think creatively; the description of the study included instructions intended to emphasize that people ought to try to be creative. For example, part of the description was: "Our study today is about how people think creatively, like how people come up with original, innovative ideas. Everyone can think creatively, and we'd like to learn more about how people do it. So today people will work on a few different creativity tasks, which look at how people think creatively."

We think that it is essential to instruct people to try to be creative, for three reasons. First, instructing people to be creative increases the creativity of their responses (e.g., Christensen et al., 1957; Runco, Illies, & Eisenman, 2005), which will raise the ceiling of creativity in the sample and hopefully expand the variance in creativity scores. Second, instructing people to be creative makes the scores more valid indicators of individual differences. Harrington (1975) for example, showed that "be creative" instructions enhanced the covariance between divergent thinking scores and measures of personality (see also Katz & Poag, 1979). And third, creativity scores are ambiguous when people are not trying to be creative. Someone can achieve a low score by having a genuinely low level of creativity or by failing to realize that the study is about creativity.

We administered three divergent thinking tasks: an unusual uses task, an instances task, and a consequences task. For the *unusual uses task*, we instructed people to generate creative uses for a brick. The experimenter's instructions emphasized that the task was about creative uses:

> For this task, you should write down all of the original and creative uses for a brick that you can think of. Certainly there are common, unoriginal ways to use a brick; for this task, write down all of the unusual, creative, and uncommon uses you can think of. You'll have three minutes. Any questions?

After three minutes, the experimenter instructed everyone to stop writing and to evaluate their responses. They were told to "pick which two are your most creative ideas. Just circle the two that you think are your best." People could take as much time as they wished to pick their top two, but they took only a few moments.

For the *instances task*, we instructed people to generate creative instances of things that are round:

> For this task, you should write down all of the original and creative instances of things that are round that you can think of. Certainly there are some obvious things that are round; for this task, write down all of the unusual, creative, and uncommon instances of things that are round. You'll have three minutes. Any questions?

After the task, people circled their two most creative responses.

For the *consequences task*, people had to generate creative consequences for a hypothetical scenario: what would happen if people no longer needed to sleep. As before, we instructed them to generate creative consequences:

> For this task, imagine that people no longer needed to sleep. What would happen as a consequence? Write down all of the original, creative consequences of people no longer needing to sleep. You'll have three minutes. Any questions?

People circled their two most creative responses after the task.

Scoring the Responses

The participants in the study generated 1,596 responses. Each response was typed into a spreadsheet and then sorted alphabetically within each task. (Spelling errors were silently corrected prior to rating.) This method ensured that the raters were blind to several factors that could bias their ratings: (1) the person's handwriting; (2) whether the person circled a response as a top 2 response; (3) the response's serial position in the set; (4) the total number of responses in the set; and (5) the preceding and following responses. Three raters evaluated each response to each task. The raters read all of the responses prior to scoring them, and they scored the responses separately from the other raters. Each response received a rating on a 1 to 5 scale (1 = *not at all creative*, 5 = *highly creative*).

The scoring criteria were adopted from Wilson, Guilford, and Christensen's (1953) classic article on individual differences in originality. In their model, creative responses are uncommon, remote, and clever. In support of their model, they found that tasks designed to measure uncommonness, remoteness of association, and cleverness loaded on a single originality factor. The instructions given to the raters are shown in Appendix 1. The raters were told to consider all three dimensions when making their ratings,

and they were told (following Guilford)that strength in one facet can balance weakness in another facet. Two specific additional criteriawere used, following recommendations by Harrington et al. (1983). For the uses task, the raterswere told to give lower scores to actual uses for bricks (e.g., making a wall or a fireplace); for theinstances task, the raters were told to give lower scores to round objects visible in the researchroom (e.g., bottles of water, pens and pencils).

Forming the Creativity Indexes

We calculated three creativity indexes for analysis.

Average creativity. The first and most straightforward index is the average rating of all ofthe responses. For this index, the person's ratings were summed and then divided by the numberof responses. This index takes into account the entire set of responses: someone with 3 creativeresponses will have a higher average than someone with 3 creative responses and 5 uncreativeresponses. The *Average creativity index* thus imposes a penalty for generating many uncreativeresponses.

Top 2 creativity. The second index averaged the ratings of the responses that people choseas their two best responses. Unlike Average scoring, *Top 2 scoring* constrains the number ofresponses that are assessed and thus omits some responses. For this index, someone with 3creative responses (2 picked as the top 2) will have similar scores as someone with 2 creativeresponses (both picked as the top 2) and 5 uncreative responses. Because it includes only the best

scores, as decided by the respondent, the Top 2 index doesn't penalize people for generatingmany uncreative responses. (If a person had only one response, the value for that response wasused. If a person had no responses, the data were labeled as missing.)

Uniqueness. The third index was the classic 0/1 *uniqueness index* developed by Wallachand Kogan (1965). People received a 1 for each response that was unique in the sample and a 0for each response that was given by at least one other person. The unique responses weresummed to create scores for each person.

An Overview the Generalizability Analyses

G-theory allows the researcher to define a universe of admissible observations bydetermining facets. Facets are measurement conditions that are seen as similar andinterchangeable. For Average scoring and Top

2 scoring, we included the object of measurement (the examinees, which are not considered a source of error) and two facets of measurement error: the rater facet and the task facet. G-theory can treat facets as random or fixed. In the current design, raters are considered random, tasks were treated as random initially (but the results suggested changing them to a fixed facet[3]), and scoring type was included as fixed (treating each scoring type separately). For uniqueness scoring, we included the object of measurement and one facet of measurement error: the task facet.

The difference between random and fixed lies in the idea of interchanging measurement conditions. For example, raters are considered interchangeable, which means that the score evaluated by one rater should be consistent with a score given by another rater. In other words, theoretically there is an infinite pool of raters, and if we randomly draw a set of three raters, their scoring of a divergent thinking task should produce roughly the same observed score as another random set of three raters evaluating the same task by the same examinee. In the case of a fixed facet, here the scoring type, there is not an infinite universe of scoring types. Hence, we do not randomly sample scoring types or see them as interchangeable. These conceptualizations are similar to how factors would be defined in ANOVA. In terms of interpretation, the score on a divergent thinking task reached by one scoring type (e.g., Average scoring) does not provide information about the score from another scoring type (e.g., uniqueness). We consider scoring type as a fixed facet because, conceptually, we do not view the scoring methods as interchangeable. For the univariate analysis, GENOVA (Crick & Brennan, 1983) was used to obtain variance component estimates and generalizability and Φ coefficients.

Results

Descriptive Statistics

Table 1 shows the descriptive statistics for the creativity scores for each task. These scores are averaged across the 3 raters.

Generalizability and Dependability

Our generalizability analyses are broken into three steps. First, we present the results for the person-by-rater design (p x r design) for Average scoring and Top 2 scoring. This design assumes that tasks are fixed effects, so each task is analyzed separately. Second, we present the results for the person-by-task design (p x t

design) for uniqueness scoring. Sample sizes differedslightly due to missing data, given that GENOVA requires listwise deletion of missing data.*Generalizability Study for p x r Design for Average Scoring and Top 2 Scoring*

For our primary analysis, we used a p x r design: task was estimated as a fixed facet. Theresults for Average scoring for each task are shown in Table 2. The unusual uses task and the

instances task performed equally well, but the consequences task stuck out. The variancesexplained by real performance differences between examinees were high (62.6% for the unusualuses task and 63.9% for the instances task). The raters accounted for some of the variance (10.4%for the unusual uses task and 12.4% for the instances task). This result indicates that raters wereslightly inconsistent in their ratings across examinees—some raters are consistently morestringent than others—but the variance due to raters appears modest in light of the variance dueto examinees. The confounded components of person-by rater interaction and random errorwere substantial (27% for the unusual uses task and 23.7% for the instances task).

For the consequences task, we had less variance associated with people (34%) and muchmore variance introduced by raters (37.4%). This was a large variance component: the ratersbehaved inconsistently across people for this task. If more variance in scores is due to the ratersthan to the test takers, then the task is a poor measure of the test takers' performance. As we'llshow later, researchers would need a lot of raters to obtain a dependable divergent thinking scoreon the consequences task.

For Top 2 scoring (see Table 3), the pattern for all three tasks mirrored the pattern forAverage scoring. Overall, this scoring approach was slightly less dependable. The varianceassociated with people was smaller than in Average scoring (56.2% for the unusual uses task and50.0% for instances task), although it was at least half of the variance in each case. The randomerror with the person-by-rater interaction increased (40.6% for the unusual uses task and 41.4%for the instances task). The variance accounted for by rater inconsistencies was smaller (3.2% forthe unusual uses task and 8.6% for the instances task), but that shouldn't be considered an improvement in light of the other components in the model. The consequences task, as with

Average scoring, performed the worst: the variances accounted for by people (34.9%), raters(28.3%), and error with person-by-rater interaction (36.8) were about equal.

Dependability Coefficients for p x R Design for Average Scoring and Top 2 Scoring

The decision study forecasts what dependability scores a researcher could expect under variations of the design. Table 4 shows G-coefficients (for relative decisions) and Φ coefficients (for absolute decisions). Like Cronbach's alpha coefficients, they range from 0 to 1, and higher values reflect more dependable scores. As with alpha, .80 can serve as an informal threshold for reliable scores (DeVellis, 2003). By means of D-study analysis, we also can estimate dependability estimates associated with other measurement designs. Researchers planning divergent thinking research can use these estimates to choose which tasks to use and how many raters to train.

These dependability estimates show several trends. First, Average scoring had higher coefficients than Top 2 scoring. Table 4 shows that, on the whole, Average scoring produced more dependable scores for all three tasks and for all numbers of raters. Second, the unusual uses task and the instance task produced similarly dependable scores, but the consequences task produced less dependable scores. Third, the effect of adding raters on dependability diminishes quickly. In general, increasing raters from 1 to 2 has a large effect on dependability, and increasing raters from 2 to 3 has an appreciable effect. The gain from increasing raters to 4 is small, and little is gained from going from 4 to 5 raters. Finally, as expected, the Φ coefficients were consistently lower than the G-coefficients.

Uniqueness Scoring

Uniqueness scoring has a task facet but no rater facet: a single person coded whether or not a response was unique within the pool of responses. As Table 5 shows, examinees accounted for 15.9% of the variance, tasks accounted for 28.5% of the variance, and the interaction of person and task including the random error accounted for 55.6% of the variance. This result indicates that tasks differed in terms of difficulty. The interaction of person and task including the random error explained the largest amount of variance. We only can speculate about the reasons because this variance component is confounded. This variance may be attributable to people performing differently across tasks, to random error, or to both. Overall, users of this scoring technique can't expect dependable scores. The dependability coefficients for the uniqueness scoring in the 1-facet design (p x t) were poor (see Table 6). To get dependable scores, researchers would need 15 tasks to get .81 for relative decisions and 20 tasks to get .79 for absolute decisions.

Was Creativity Distinct From Fluency?

The confounding of creativity scores and fluency scores has plagued divergent thinking research for several decades. Table 7 shows the relationships of the creativity scores with fluency (i.e., the number of responses generated on the task). The two subjective scoring methods performed well, but the uniqueness scoring method showed the usual high correlations with fluency. The Pearson correlations between creativity and fluency ranged from -.23 to -.05 for Average scoring, from -.18 to .09 for Top 2 scoring, and from .35 to .67 for uniqueness scoring. The Average and Top 2 indexes thus apparently avoid the fluency confound that pervades research with the uniqueness index. For the Average and Top 2 scores, people with creative scores were not necessarily people who generated a lot of responses. The small negative coefficients, in fact, indicate that generating a lot of responses predicted somewhat less creative responses.

We should point out that fluency scores have a different meaning in our study than in studies that did not instruct people to be creative (e.g., Wallach & Kogan, 1965). Telling people be creative causes fewer responses (Christensen et al., 1957; Harrington, 1975), probably because people use quality-over-quantity strategies instead of mere-quantity strategies. Thus, the scores represent the number of responses people generated while trying to generate creative responses, not the number of responses people could generate when trying to generate as many as possible. We suspect that both the average level of fluency and the variance in fluency is lower in our study, which would deflate correlations between fluency and other variables.

Discussion

Study 1 explored the dependability of subjective ratings of divergent thinking tasks. Both the Average scoring and the Top 2 scoring performed well for most of the tasks. For the unusual uses and instances tasks, both scoring methods yielded dependable scores ($G > .80$) with two or three raters. For the consequences task, participants and raters contributed equal variance to the scores; this task would require four or five raters for dependable scores. We compared these

scores to the Wallach and Kogan (1965) uniqueness scoring. According to our analyses, uniqueness scoring requires many tasks (around 15) to reach a dependability of .80. Moreover, only the uniqueness scoring showed appreciable relationships with fluency—the subjective scoring methods yielded scores that were essentially unrelated to fluency.

We should emphasize that our findings—both the variance decomposition and the coefficients of dependability—are based on a design with 73 examinees, 3 raters, and 3 tasks. As in classical test theory, the results are sample-dependent. Replications would provide information on how much these estimates vary from sample to sample. If researchers have the resources, then it would be helpful to run a G-study and D-study during a piloting phase to get information on how many raters are needed to get dependable scores. By applying G-studies and D-studies, the precision of measurement can be greatly increased—researchers can understand and thus reduce the sources of error in their tools for measuring creativity.

Study 2: Validity of Average and Top 2 Scoring

Study 1 suggested that the scores from both subjective scoring methods had good reliability. What about validity? Study 2 sought evidence to support our interpretation of the divergent thinking scores. To appraise validity, we examined the relationship between divergent thinking scores and broad dimensions of personality. Creativity research has a long tradition of research on relationships between divergent thinking and individual differences (for reviews, see Joy, 2004; Runco, 2007; Sawyer, 2006; Weisberg, 2006), so personality provides a meaningful context for appraising whether the divergent thinking scores behave as they should.

The five-factor model of personality is a natural place to start when exploring personality and creativity. Five-factor theories propose that personality structure can be captured with five broad factors: neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (McCrae & Costa, 1999). In creativity research, openness to experience is the most widely-studied of the five factors. If any personality trait is a "general factor" in creativity, openness to experience would be it. First, openness is associated with divergent thinking. Past research with the Guilford tasks (McCrae, 1987) and the Torrance verbal tasks (Carson, Peterson, & Higgins, 2005; King, Walker, & Broyles, 1996) has found medium-sized effects between openness and divergent thinking (r = .30, Carson et al., 2005; r = .34, McCrae, 1987; r = .38, King et al., 1996). Second, openness is associated with other aspects of a creative personality,

such as viewing oneself as a creative person and valuing originality (Joy, 2004; Kaufman & Baer, 2004). Finally, openness is associated with creative accomplishment in diverse domains, such as science and the arts (Feist, 1998, 2006).

The five factors form two higher-order factors (DeYoung, 2006; Digman, 1997). *Plasticity*—composed of openness to experience and extraversion—reflects a tendency toward variable, flexible behavior. It captures the novelty-seeking and unconventional qualities of openness and the impulsive and energetic qualities of extraversion. *Stability*—composed of neuroticism (reversed), agreeableness, and conscientiousness—reflects a tendency toward controlled, organized, regulated behavior. It captures the stable moods and self-perceptions of emotional stability (the other pole of neuroticism); the empathetic, friendly, and accommodating qualities of agreeableness; and the self-control of conscientiousness. Plasticity and stability resemble other well-known dichotomies in personality psychology (Digman, 1997), such impulsiveness versus constraint (Carver, 2005) and ego-resiliency versus ego-control (Letzring, Block, & Funder, 2005). Research has not yet examined relations between creativity and the higher-order factors, which for convenience we'll call the Huge 2.

We assessed commitment to a college major in the arts as a predictor of divergent thinking. Our sample, which is primarily first-year college students, is too young to examine the relationship between creative accomplishments across the lifespan and divergent thinking (cf. Plucker, 1999). But we can measure whether people have committed to an artistic occupation, thus capturing indirectly people's creative interests (Feinstein, 2006) and providing concurrent evidence for validity. People pursuing arts majors—majors devoted to the fine arts, performing arts, or decorative arts—have chosen to devote their college years to receiving training in an artistic field, and training is necessary for later creative accomplishment (Sawyer, 2006). Variability in college majors can thus represent the creativity of people's occupational and lifespan goals.

Method

Participants and Design

A total of 242 students enrolled in General Psychology at UNCG participated in the "Creativity and Cognition" project and received credit toward a research option. We excluded 16 (6.6%) people who showed limited English proficiency, who had extensive missing data, or who gave capricious responses to the questionnaire (e.g., circling the midpoint for most items). This left us with a final sample of 226 people (178 women, 48 men). According to self-reported ethnic identification, the sample was 63% European American, 27% African American, and 10% other ethnic groups. Most people (82%) were 18 or 19 years old. The most

common college majors were nursing (31%), undecided (11%), and biology (5%); fewer than 3% were psychology majors.

Procedure

People participated in 90-minute sessions in groups of 1 to 13. After providing informed consent, people learned that the study was about the psychology of creativity. The experimenter explained that the researchers were interested in how creativity related to various aspects of personality, attitudes, and thinking styles. People completed several creativity tasks, cognitive tasks, and measures of personality; we present the findings for personality and the unusual-uses tasks here.

Divergent Thinking Tasks

The experiment began with the divergent thinking tasks. Study 1 found that the unusual- uses task had the highest reliability, so in Study 2 we measured individual differences in divergent thinking with two unusual uses tasks: uses for a brick and uses for a knife. We used the same instructions and procedure as in Study 1: we instructed people to be creative, people had three minutes per task, and they circled their top two responses after each task.

Big 5 Scales

We measured the Big 5 domains of personality with three scales. For the first scale, we used Costa and McCrae's (1992) 60-item Five Factor Inventory, which measures each domain with 12 items. People responded to complete sentences (e.g., the Openness item "Sometimes when I am reading poetry or looking at a work of art, I feel a chill or wave of excitement") on a 5-point scale (1 = *strongly disagree*, 5 = *strongly agree*). For the second scale, we formed a 50-item Big 5 scale from the Big 5 items in Goldberg's public-domain International Personality Item Pool (IPIP; Goldberg et al., 2006). Each domain was measured by 10 items. People rated how well sentence fragments described them (e.g., the Openness item "Am full of ideas") on a 5-point scale (1 = *very inaccurate description of me*, 5 = *very accurate description of me*). For the third scale, we used Gosling, Rentfrow, and Swann's (2003) 10-item brief Big 5 scale, which measures each domain with two items. Each item has two adjectives, and people rated how well the adjective pair described them (e.g., the Openness

item "Open to new experiences, complex") on a 5-point scale (1 = *very inaccurate description of me*, 5 = *very accurate description of me*).

Arts Major

On a demographic page, people listed their major in college. We classified each person's college major as either an arts major (1 point) or a conventional major (0 points). All majors, concentrations, and certification programs (called simply "majors") concerned with the fine arts, performing arts, and decorative arts were classified as arts majors. When necessary, we consulted UNCG's Undergraduate Bulletin and faculty who taught in the department for more information about the major. Twenty-one people (9%) had arts majors; 205 people (91%) had conventional majors. The arts majors were acting, apparel products design, art education, art history, dance, graphic design, interior architecture, music, music education, vocal performance, fine art, studio art, theatre, and theatre education.

Results

Scoring the Divergent Thinking Tasks

People generated 3,224 responses: 1,641 for the brick task, and 1,583 for the knife task. People's Top 2 responses made up 27.5% of the brick responses (452 responses) and 28.5% of the knife responses (451 responses; one participant generated only one knife response). Three raters evaluated each response following the same instructions and methods as in Study 1. Table 8 reports the descriptive statistics for the variables. There were no missing observations.

To simplify the reporting of a large number of effects—and to recognize that small effects will be significant with a large sample—we describe our findings in terms of small (around ß = .10), medium (around ß = .30), and large (around ß = .50) effect sizes. A Web appendix, available on the first author's Internet page, provides the details (e.g., the standardized effects, unstandardized effects, intercepts, variances, standard errors, and residual variances) for each component of the 12 latent variable models reported here. For clarity, Figures 1–4 depict simplified models that omit indicators, factor covariances, and residual variances.

Model Specification and Model Fit

We estimated the latent variable models with Mplus 4.21. (The estimates were essentially identical when estimated with AMOS 6.0.) Our first step involved estimating measurement models for divergent thinking, for the Big 5 factors, and for the Huge 2 factors. To assess model fit, we considered the root mean-square error of approximation (RMSEA), the standardized root mean square residual (SRMR), the comparative fit index (CFI), the chi-square divided by the degrees of freedom (χ^2/df), and the chi-square test (χ^2). The RMSEA accounts for a model's complexity: values less than .10 indicate moderate fit, and values less than .05 indicate close fit (Browne & Cudeck, 1993). The SRMR indicates the average absolute difference between the sample correlation matrix and the correlation matrix implied by the model. Values less than .10 are good (Kline, 2005). The comparative fit index (CFI) indicates how well the fit of the predicted model improves upon the fit of a null model. CFI values greater than .95 are seen as good (Hu & Bentler, 1999). Ratios of χ^2/df less than 2 indicate good fit (Byrne, 1989, p. 55).

We modeled Divergent Thinking as a higher-order latent variable composed of two latent variables: a Brick variable and a Knife variable. The paths from the higher-order variable to the Brick and Knife variables were constrained to be equal for identification; the variance of Divergent Thinking was set to 1. The three raters' scores were the observed indicators for the Brick and Knife variables. Because the paths for the first and third raters were nearly identical for the Top 2 model, we constrained them to be equal for the Top 2 model (but not the Average model). For both models, we set the paths for the second rater to 1. The fit of this model was good for the Top 2 scores (RMSEA = .044, SRMR = .048, CFI = .99, $\chi^2/df = 1.43$, $\chi^2(10) = 14.32$, $p < .16$) and for the Average scores (RMSEA = .04, SRMR = .026, CFI = .99, $\chi^2/df = 1.36$, $\chi^2(8) = 10.88$, $p < .21$).

We modeled the Big 5 factors as five latent variables, each indicated by three scales. For each factor, we set the path to the IPIP scale to 1. The fit of the Big 5 model was not as good as the fit of the Divergent Thinking model (RMSEA = .106, SRMR = .087, CFI = .88, $\chi^2/df = 3.54$, $\chi^2(80) = 283.17$, $p < .001$). For the Huge 2 factors, the higher-order Plasticity variable was composed of the latent Openness and Extraversion variables; Plasticity's paths to these variables were constrained to be equal for identification. The higher-order Stability variable was composed of the latent Neuroticism, Agreeableness, and Conscientiousness variables. The variances of Plasticity and Stability were set

to 1. The fit of this model was about the same as the Big 5 model (RMSEA = .102, SRMR = .086, CFI = .88, Q^2/df = 3.36, $Q^2(85) = 285.73$, $p < .001$). Neither of the personality models had strong fit, but the models were retained because they represent theories of personality structure (McCrae & Costa, 1999). Because the divergent thinking model fit well, misfit in the full structural models is likely due to the personality factors, not the divergent thinking factor.

Big 5 and Divergent Thinking

We first examined how the Big 5 factors predicted divergent thinking. Figure 1 (simplified for clarity) presents the standardized path estimates. For Top 2 scores, the five factors explained 49.4% of the variance (RMSEA = .073, SRMR = .073, CFI = .89, Q^2/df = 2.21, $Q^2(175)$ = 386.65, $p < .001$). Openness (ß = .586) and Conscientiousness (ß = -.464) had large effect sizes; Agreeableness had a smaller effect size, and Extraversion and Neuroticism explained little variance. For Average scores, the five factors explained 17.2% of the variance (RMSEA = .072, SRMR = .072, CFI = .91, Q^2/df = 2.19, $Q^2(173) = 378.47$, $p < .001$). Openness (ß = .306) and Conscientiousness (ß = -.297) had medium effect sizes; Agreeableness, Extraversion, Neuroticism explained little variance. In short, the Top 2 scores appeared to have much better validity, as indexed by variance explained and by the effect sizes.

Our second set of models entered arts major as an additional predictor; Figure 2 depicts the standardized path estimates. For Top 2 scores, entering arts major increased the variance explained to 57.6% (RMSEA = .069, SRMR = .07, CFI = .90, Q^2/df = 2.08, $Q^2(190) = 396.00$, $p < .001$). Arts major had a moderate effect size (ß = .339). For Average scores, entering arts major increased the variance explained to 21.8% (RMSEA = .069, SRMR = .069, CFI = .91, Q^2/df = 2.07, $Q^2(188) = 388.79$, $p < .001$). Arts major had a smaller effect (ß = .236). As before, the Top 2 scores performed better than the Average scores: the model for Top 2 scores explained over half of the variance in divergent thinking.

Huge 2 and Divergent Thinking

What about the higher-order factors of the Big 5? Our next set of models examined how the Huge 2—Plasticity and Stability—predicted Top 2 and Average scores. As before, we first examined personality alone and then entered creative major. Figure 3 presents the standardized path estimates. For Top 2 scores, Plasticity and

Stability explained 42.1% of the variance (RMSEA = .073, SRMR = .077, CFI = .89, χ^2/df = 2.19, $\chi^2(183)$ = 401.47, $p < .001$). Plasticity had a large effect (ß = .642), and Stability had a medium effect in the other direction (ß = -.322). For Average scores, Plasticity and Stability explained 17% of the variance (RMSEA = .072, SRMR = .076, CFI = .91, χ^2/df = 2.16, $\chi^2(181)$ = 390.56, $p < .001$). Both Plasticity (ß = .388) and Stability (ß = -.247) had medium effect sizes. As before, the Top 2 scores performed better than the Average scores.

Our next analyses entered arts major as a predictor. Figure 4 presents the standardized path estimates. For Top 2 scores, entering arts major increased the variance explained to 54/0 (RMSEA = .069, SRMR = .076, CFI = .89, χ^2/df = 2.06, $\chi^2(201)$ = 414.46, $p < .001$). Arts major had a medium effect (ß = .337). For Average scores, entering arts major increased the variance explained to 22/0 (RMSEA = .068, SRMR = .075, CFI = .91, χ^2/df = 2.04, $\chi^2(199)$ = 405.93, $p < .001$). Arts major had a smaller effect (ß = .210). As before, personality and creative major collectively explained over half of the variance in Top 2 scores.

Did Fluency Confound Top 2 Scores?

Did Top 2 scores perform well by virtue of an association with fluency scores? We have argued that subjective scoring avoids the confounding of creativity with fluency, and Study 1 found small relationships between the ratings and fluency scores. The same small effects appeared here. We estimated all four Top 2 models and included a latent fluency variable. Fluency was composed of the standardized fluency scores for the Brick and Knife tasks; the paths were constrained to be equal, and the variance of fluency was set to 1. In all four models, fluency had small effects on divergent thinking: ß = .122 for the Big 5 model, ß = .103 for the Big 5 & Arts Major model, ß = .06 for the Huge 2 model, and ß = .055 for the Huge 2 & Arts Major model. It's clear, then, that fluency contributed little variance to the Top 2 scores.

Discussion

Study 2 found evidence for the validity of our subjective scoring methods. Both Top 2 scoring and Average scoring performed well, but Top 2 scoring was the clear winner. Table 9 depicts the percentage of variance explained by the latent variable models. In each case, at least twice as much variance was explained in Top 2 scores than in Average scores. For all of the models, Top 2 scores and Average scores showed the same

patterns, but the effect sizes for Top 2 scores were consistently larger. Several large effects (i.e., ß > .50) were found for Top 2 scores, but no large effects were found for Average scores. And these effects were independent of fluency scores, which contributed little to the prediction of divergent thinking. It's worth pointing out that we used only two tasks and three raters—it wasn't necessary to pool information from a lot tasks and raters to find these effects. Furthermore, the personality scores and the divergent thinking scores came from different methods: self-reports for personality, raters' judgments of performance on timed tasks for divergent thinking. Taken together, the concurrent evidence for validity is compelling enough to motivate future research.

We used personality as a context for examining validity, but the personality findings are interesting in their own right. For Top 2 scores, we found large effects of Openness to Experience on divergent thinking, and, intriguingly, Conscientiousness. High Openness predicted high creativity, but high Conscientiousness predicted low creativity. Although Openness gets the most attention, research has found strong but complex relationships between creativity and Conscientiousness. In Feist's (1998) meta-analysis, scientists were more conscientious than nonscientists, but artists were less conscientious than non-artists. Our sample of young adults can't address the role of conscientiousness in domain-specific accomplishment; this issue deserves more attention in future work.

The relations of Openness and Conscientiousness were mirrored among the Huge 2 factors. Plasticity and Stability predicted divergent thinking in opposing directions: Plasticity had a large positive effect, and stability had a medium negative effect. Plasticity's effect was larger than the effects of Openness and Extraversion, its lower-order variables. Perhaps what Openness and Extraversion share—a tendency toward approach-oriented action—is more important than their individual features. The psychology of creativity's focus on Openness may be overlooking a much stronger higher-order relationship. Finally, it is interesting that pursuing a college major in the arts consistently had a medium effect on divergent thinking. Future research is needed to unravel the intriguing meanings of this relationship, which could reflect an effect of training on divergent thinking, an effect of divergent thinking on what people choose to pursue, or a common third factor.

General Discussion

Despite its venerable history, divergent thinking has a bad reputation in sociocultural theories and cognitive theories of creativity (Sawyer, 2006; Weisberg, 2006). When faced with a body of modest effects, researchers should examine the quality of their measurement tools. With instruments that yield unreliable scores, researchers are unlikely to find large effect sizes and consistent relationships. Our two studies explored the value of two subjective-scoring methods. These studies generated a lot of information: we unpack our findings below.

Reliability

Numbers of Raters and Tasks

The reliability of divergent thinking scores is due in part to the number of tasks that people complete. Thus far, there is no empirical guidance for how many tasks is sufficient or appropriate. A glance at research shows incredible variance: studies have administered 1 task (Silvia & Phillips, 2004), 3 tasks (Hocevar, 1979a), 4 tasks (Carson et al., 2005), 9 tasks (Katz & Poag, 1979), and 15 tasks (Runco, 1986). Wallach and Kogan (1965), in their classic research, set the record—they administered 39 divergent thinking tasks. One task is probably not enough for dependable scores; 39 is probably excessive. It's hard to tell, based on intuition, how many tasks ought to be used. And when using subjective ratings, researchers are faced with deciding how many raters are necessary to obtain reliable scores. To date, research on divergent thinking has used a wide range of raters, such as 1 rater (Wilson et al., 1953), 2 raters (Christensen et al., 1957; Silvia & Phillips, 2004), 3 raters (Grohman et al., 2006; Harrington, 1975; Mouchiroud & Lubart, 2001), and 4 raters (Hocevar, 1979b). Research using the consensual assessment technique has a wide range as well, such as 5 raters (Carson et al., 2005), 13 raters (Kaufman et al., 2005), and 20 raters (Amabile, 1982). One rater is clearly not enough; 20 seems like overkill.

Generalizability theory can provide practical guidelines for research by estimating how tasks and raters contribute to reliability. In Study 1, we used three tasks—an unusual uses task, and instances task, and a consequences task—that are typical of the kinds of verbal tasks used in divergent thinking research (Runco, 2007) and in creativity testing (Torrance, 2008). We found that the unusual uses task and the instances task functioned similarly, but the consequences task deviated from both. Under Average scoring, to get a dependable divergent

thinking score of above .80 for relative decisions, we would need two raters for the unusual uses task (.82), two raters for the instances task (.84), but three raters for the consequences task (.83). The Top 2 scoring functioned less well overall, but the pattern for the three tasks was similar. Under Top 2 scoring, to reach a dependability of above .80, we would need three raters for the unusual uses task (.81), four raters for the instances task (.83), but five raters for the consequences task (.83).

It's noteworthy, we think, that consequences tasks have performed badly in other studies. In Harrington's (1975) experiment, consequences tasks were administered but not analyzed because the raters were unable to achieve reliable scores (Harrington, 1975, p. 440). In our own study, the participants and the raters accounted for similar amounts of variance on the consequences task. Researchers can enhance the dependability of a consequences task by adding more raters, but they might prefer to use other, more efficient tasks instead.

Which Scoring Methods Performed Best?

We have proposed that creativity researchers should use subjective scoring of divergent thinking tasks. Study 1 compared the reliability of three scoring methods: Average scoring, Top 2 scoring, and Wallach and Kogan's (1965) uniqueness scoring. To start with the weakest method, we found that the uniqueness index functioned badly. For a dependability level of .80, researchers would need 15 tasks (see Table 6). With fewer tasks, the uniqueness index will provide undependable scores. For example, imagine a study that administers 4 tasks —a typical amount for divergent thinking research—and uses uniqueness scoring. According to Table 6, the scores would have a dependability level of .53. This value is bad: it wouldn't be acceptable for a measure of attitudes, personality, or individual differences (DeVellis, 2003). In light of the poor dependability of the uniqueness index, it isn't surprising that divergent thinking research rarely finds large effects.

The two scoring methods based on subjective ratings, in contrast, performed well. Both Average scoring and Top 2 scoring produced dependable scores; researchers can expect dependability levels of .80 with 2 or 3 raters (Table 4). Researchers should keep in mind that these dependability estimates are for studies that use our administration and scoring guidelines, which we have described in detail in the text and in Appendix 1.

Were Subjective Ratings Eccentric and Idiosyncratic?

Many researchers are skeptical of subjective ratings, believing them to be eccentric and idiosyncratic. But whether raters agree is an empirical matter, and we can easily evaluate how consistently raters judged the divergent thinking tasks. Overall, Study 1 found good levels of agreement among the raters for the unusual uses and instances tasks. For Average scoring, raters accounted for around 10–12% of the total variance; for Top 2 scoring, raters accounted for 4–8% of the total variance. And in each case, the variance due to performance differences between the participants was many times greater than the variance due to the raters. The pattern of variance—participants accounted for 50–60% of the variance and raters accounted for 4–12% of the variance—should allay concerns about whether these tasks are merely capturing willy-nilly differences between raters.

Another way to understand the good level of agreement between raters is to examine how many raters are needed to have dependable scores. The dependability estimates (see Table 4) indicate that the gain in dependability diminishes after 3 or 4 raters. All of the G and Φ coefficients are over .80 for 4 raters, so researchers would rarely need to recruit and train more than 4 raters. These values are practical for people working in the trenches of creativity research.

Validity

Predicting Divergent Thinking Scores

According to most theories of validity, evidence for validity comes from establishing relationships between the construct of interest and other constructs (Cronbach & Meehl, 1955; Messick, 1995). Validity is never established definitively, but our first study of validity offered support for our assessment method. The Big Five factors and the creativity of people's college majors collectively explained 57% of the variance in the Top 2 scores of two unusual uses tasks (see Table 9). The other Top 2 models fared well, too, explaining at least 42% of the variance. Concurrent evidence for validity came from relationships with openness to experience, conscientiousness, and the creativity of people's college majors. To expand the evidence for validity, future research should explore other constructs—such as individual differences in cognition, attitudes, and creative accomplishment—and other research designs, such as longitudinal designs.

It's interesting, we think, that the Top 2 scores had stronger evidence for validity than the Average scores. People's best responses, as defined by the two they chose as their most creative, carry more information than all of their responses. This shouldn't be too surprising: most participants give at least a few uncreative responses, so the uncreative responses are more or less constant across people. The responses that discriminate between people are the best responses. Because the Top 2 scoring method evaluated only the best two responses, it omits responses that are less informative.

Was Creativity Confounded With Fluency?

Uniqueness scoring confounds the number of responses with the quality of responses (Hocevar, 1979a, 1979b). In Study 1, we found the usual high positive correlations between fluency scores and uniqueness scores. But fluency was essentially unrelated to the Average scores and Top 2 scores. A few of the correlations were modestly negative, indicating that people with creative responses tended to produce fewer responses overall. In Study 2, a latent fluency variable was essentially unrelated (a range of ß = .122 to ß = .055) to latent Top 2 scores. Past research has suggested many ways of handling the fluency confound, but these methods have generally not performed well psychometrically (see Michael & Wright, 1989). The subjective scoring methods, in contrast, sharply separate creativity from fluency and produce dependable, valid scores.

Summary of Recommendations for Researchers

Researchers can use Tables 4 and 6 to estimate the dependability of their measurement when designing experiments. Based on our two studies and other findings, we offer these take- home recommendations for researchers interested in using our approach to assessment and scoring:

1. Researchers should instruct participants to be creative. Several studies have shown that "creativity instructions" enhance the validity of divergent thinking scores (Harrington, 1975).
2. The traditional Wallach and Kogan (1965) uniqueness index fared badly: it will give dependable scores only with many tasks. To achieve a dependability level of around .80, researchers will need 15 tasks (see Table 6).
3. The unusual uses and instances tasks performed better than the consequences task in Study 1, and two unusual uses tasks performed well in Study 2. Unless researchers are specifically interested in consequences tasks, they probably ought to pick other classes of divergent thinking tasks.

4. Concerning reliability, both Average scoring and Top 2 scoring worked well; Averagescoring was slightly more dependable. To achieve a dependability level of around .80 for thesescoring methods, researchers will need 2 raters for Average scoring and 3 raters for Top 2 scoring(see Table 6). Researchers will rarely need more than 4 raters. We recommend that researcherscollect and analyze both kinds of scores. Top 2 scores are a subset of Average scores, so theyrequire little extra effort.
5. Concerning validity, Top 2 scores were the clear winner. Top 2 scores had consistentlylarger effects than Average scores, and the models explained at least twice as much variance inTop 2 scores than in Average scores. Judging people on their best responses appears to be an effective way of assessing individual differences in divergent thinking.

Conclusion and Invitation

The psychology of creativity ought to be open to innovative approaches to assessment. We can guarantee that our Top 2 scoring method is not the best of all possible methods, but ourresearch has shown that it performs well: the evidence for reliability is good, and we explained ahuge amount of variance in divergent thinking. We encourage researchers to continue to developnew and refined approaches to assessment. To accelerate the development of better methods, wehave archived the data from Studies 1 and 2. We invite researchers to use these data as benchmarks for comparing new approaches. Researchers can apply new scoring methods to theresponses and then directly compare which method performs better. Is our method better than the typical Torrance scores of fluency, originality, and flexibility (Torrance, 2008)? Is a singlesnapshot score—one rating given to the entire set of responses (Mouchiroud & Lubart, 2001)—better than ratings of each response? Is it better to use the participant's chosen top tworesponses, or are the two responses that received the highest ratings better? Do the Big 5 factorsexplain more variance in a new scoring method than in our Top 2 method? Reliability and validity are empirical questions; we're curious to see the answers.

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. (1996). *Creativity in context.* Boulder, CO: Westview.

Baer, J., & Kaufman, J. C. (2005). Whence creativity? Overlapping and dual-aspect skills and traits. In J. C. Kaufman & J. Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 313–320). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30,* 149–165.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage. Byrne, B. M. (1989). *A primer of LISREL*. New York: Springer.

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.

Carver, C. S. (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. *Personality and Social Psychology Review, 9,* 312–333

Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology, 53,* 82–88.

Clark, P. M., & Mirels, H. L. (1970). Fluency as a pervasive element in the measurement of creativity. *Journal of Educational Measurement, 7,* 83–86.

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.

Crick, G. E., & Brennan, R. L. (1983). *GENOVA* [Computer software]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

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.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Newbury Park, CA: Sage.

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.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of Openness/Intellect: Neuropsychological correlates of the fifth factor of personality. *Journal of Personality, 73,* 825–858.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73,* 1246–1256.

Dixon, J. (1979). Quality versus quantity: The need to control for the fluency factor in originality scores from the Torrance Tests. *Journal for the Education of the Gifted, 2,* 70–79. Feinstein, J. S. (2006). *The nature of creative development.* Stanford, CA: Stanford University Press.

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.

Getzels, J. W., & Jackson, P. W. (1962). *Creativity and intelligence: Explorations with gifted students.* New York: Wiley.

Gibbons, J. D. (1993). *Nonparametric measures of association.* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07–091). Newbury Park, CA: Sage.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (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., & K#usak, M. (2006). Divergent thinking and evaluation skills: Do they always go together? *Journal of Creative Behavior, 40,* 125–145.

Guilford, J. P. (1950). Creativity. *American Psychologist, 5,* 444–454.

Harrington, D. M. (1975). Effects of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores. *Journal of Personality, 43,* 434–454.

Harrington, D. M., Block, J., & Block, J. H. (1983). Predicting creativity in preadolescence from divergent thinking in early childhood. *Journal of Personality and Social Psychology, 45,*609–623.

Hocevar, D. (1979a). A comparison of statistical infrequency and subjective judgment as criteria in the measurement of originality. *Journal of Personality Assessment, 43,* 297–299. Hocevar, D. (1979b). Ideational fluency as a confounding factor in the measurement of originality. *Journal of Educational Psychology, 71,* 191–196.

Hocevar, D., & Michael, W. B. (1979). The effects of scoring formulas on the discriminant validity of tests of divergent thinking. *Educational and Psychological Measurement, 39,*917–921.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55. Joy, S. (2004). Innovation motivation: The need to be different. *Creativity Research Journal, 16,* 313–330.

Katz, A. N., & Poag, J. R. (1979). Sex differences in instructions to "be creative" on divergent and nondivergent test scores. *Journal of Personality, 47,* 518–530.

Kaufman, J. C., & Baer, J. (2004). Sure, I'm creative—but not in mathematics! Self-reported creativity in diverse domains. *Empirical Studies of the Arts, 22,* 143–155.

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 reliability. *Thinking Skills and Creativity.*

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal, 18,* 3–14.

King, L. A., Walker, L. M., & Broyles, S. J. (1996). Creativity and the five-factor model. *Journal of Research in Personality, 30,* 189–203.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music, 35,* 144–168.

Letzring, T. D., Block, J., & Funder, D. C. (2005). Ego-control and ego-resiliency: Generalization of self-report scales based on personality descriptions from acquaintances, clinicians, and the self. *Journal of Research in Personality, 39,* 395–422.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52,* 1258–1265.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality* (2nd ed., pp. 139–153). New York: Guilford. 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.

Michael, W. B., & Wright, C. R. (1989). Psychometric issues in the assessment of creativity. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Handbook of creativity* (pp. 33–52). New York: Plenum.

Milgram, R. M., & Milgram, N. A. (1976). Creative thinking and creative performance in Israeli students. *Journal of Educational Psychology, 68,* 255–259.

Moran, J. D., Milgram, R. M., Sawyers, J. K., & Fu, V. R. (1983). Original thinking in preschool children. *Child Development, 54,* 921–926.

Mouchiroud, C., & Lubart, T. (2001). Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. *Journal of Genetic Psychology, 162,* 382–401.

Plucker, J. A. (1999). Is the proof in the pudding? Reanalyses of Torrance's (1958 to present) longitudinal data. *Creativity Research Journal, 12,* 103–114.

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: Lawrence Erlbaum Associates.

Plucker, J. A., & Renzulli, J. S. (1999). Psychometric approaches to the study of human creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 35–61). New York: Cambridge University Press.

Runco, M. A. (1986). Maximal performance on divergent thinking tests by gifted, talented, and nongifted students. *Psychology in the Schools, 23,* 308–315.

Runco, M. A. (2007). *Creativity*. Amsterdam: Elsevier.

Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *Journal of Creative Behavior, 39,* 137–148.

Runco, M. A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement, 52,* 213–221.

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment, 5,* 149–156.

Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation.* New York: Oxford University Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Silvia, P. J. (in press). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). *Creativity Research Journal.*

Silvia, P. J., & Phillips, A. G. (2004). Self-awareness, self-evaluation, and creativity. *Personality and Social Psychology Bulletin, 30,* 1009–1017.

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.Speedie, S. M., Asher, J. W., & Treffinger, D. J. (1971). Comment on "Fluency as a pervasive element in the measurement of creativity." *Journal of Educational Measurement, 8,* 125–126.

Torrance, E. P. (1967). The Minnesota studies on creative behavior. *Journal of Creative Behavior, 1,* 137–154.

Torrance, E. P. (2008). *Torrance Tests of Creative Thinking: Norms-technical manual, verbal formsA 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.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures:Nonreactive research in the social sciences.* Oxford, UK: Rand McNally.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science,invention, and the arts.* Hoboken, NJ: Wiley.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individualdifferences in originality. *Psychological Bulletin, 50,* 362–370.

Author Note

Paul J. Silvia, Christopher M. Barona, Joshua T. Cram, Karl I. Hess, Jenna L. Martinez, and Crystal A. Richard, Department of Psychology, University of North Carolina at Greensboro; Beate P. Winterstein and John T. Willse, Department of Educational Research Methodology, University of North Carolina at Greensboro.

This research was presented at the 2007 meeting of the Midwestern Psychological Association. We thank Mike Kane and Tom Kwapil for their comments on these studies. The last five authors contributed equally and are listed alphabetically. The first author's Web page (at the time of publication: http://www.uncg.edu/~p_silvia) contains the two Web appendixes mentioned in the article and, for researchers interested in recoding or reanalyzing the data, Study 2's data files and input files.

Correspondence should be addressed to Paul J. Silvia, Department of Psychology, P. O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC, 27402-6170. Electronic mail can be sent to p_silvia@uncg.edu. Phone: (336) 256-0007; Fax (336) 334-5066.

Footnotes

1. Many of the variations of uniqueness scoring cannot overcome this problem. For example, weighting each response by its frequency of occurrence (e.g., Wilson et al., 1953; Runco et al., 1987) doesn't circumvent that large-sample penalty. The base rate of uniqueness still declines with a large sample, thereby raising the criterion needed to receive a high weight. Likewise, the probability of a response within a high percentile (e.g., 95th percentile; Milgram & Milgram, 1976) declines as the sample size increases. For example, a response has a higher chance of falling above the 95th percentile in a sample of 50 responses than in a sample of 1000 responses. By giving points for not-common responses, the TTCT avoids the base-rate problem. The confounding of fluency and originality, however, is still severe for the verbal TTCT (Torrance, 2008).

2. Cronbach et al. (1972, p. 15) described it best:

The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. *The question of "reliability" thus resolves into a question of accuracy of generalization, or "generalizability."*

3. Our initial analyses treated tasks as a random facet, but the patterns of variance suggested that tasks ought to be treated as fixed. Because we do not have enough tasks for a convincing test of whether the tasks are fixed versus random, we present only the fixed-facet analyses here. Readers interested in the full analyses can download them from the first author's Web page.

Table 1

Descriptive Statistics and Correlations: Study 1

	M	*SD*	*n*	*Median*	*Skew (SE)*	*Kurtosis (SE)*	*Min/Max*	1	2	3	4	5	6	7	8	9
1. Unusual Uses: Average Creativity	1.447	0.371	76	1.34	1.605 (.276)	4.214 (.545)	1 / 3.08	1								
2. Unusual Uses: Top 2	1.706	0.636	76	1.5	.737 (.276)	-.253 (.545)	1 / 3.33	.79	1							
3. Unusual Uses: Uniqueness	1.701	1.598	77	1	.903 (.274)	.210 (.541)	0 / 6	.59	.49	1						
4. Instances: Average Creativity	1.557	0.394	75	1.44	1.292 (.277)	1.417 (.548)	1 / 2.83	.15	.29	.12	1					
5. Instances: Top 2	1.649	0.516	76	1.58	.868 (.276)	.339 (.545)	1 / 3	.18	.29	.11	.73	1				
6. Instances: Uniqueness	3.455	2.468	77	3	.767 (.274)	.290 (.541)	0 / 11	.22	.38	.29	.09	.31	1			
7. Consequences: Average Creativity	1.614	0.375	76	1.57	.480 (.276)	-.412 (.545)	1 / 2.53	.41	.29	.42	.01	.07	.28	1		
8. Consequences: Top 2	1.755	0.521	74	1.67	.448 (.279)	-.649 (.552)	1 / 3	.38	.27	.31	.17	.17	.15	.78	1	
9. Consequences: Uniqueness	1.157	1.307	76	1	1.429 (.276)	2.221 (.545)	0 / 6	.25	.31	.30	.16	.13	.11	.22	.17	1

Note. Response scales for Average Creativity and Top 2 Creativity could range from 1 to 5; the scores are averaged across the 3 raters.

Table 2

*Estimated Variance Components, Standard Error (*SE*), and Percentage of Variance Accounted for by Effects (Percent) for p x r for Average Scoring*

	Unusual Uses Task (p x r)			Instances Task (p x r)			Consequences Task (p x r)		
Effects	Variance	*SE*	Percent	Variance	*SE*	Percent	Variance	*SE*	Percent
p	0.120	0.023	62.6	0.140	0.026	63.9	0.110	0.023	34.0
r	0.020	0.015	10.4	0.027	0.020	12.4	0.121	0.086	37.4
p x r, e	0.052	0.006	27.0	0.052	0.006	23.7	0.093	0.011	28.6

Note. $n = 73$.

Table 3

Estimated Variance Components, Standard Error (SE)*, and Percentage of Variance Accounted for by Effects (Percent) for p x r for Top 2 Scoring*

	Unusual Uses Task (p x r)			Instances Task (p x r)			Consequences Task (p x r)		
Effects	Variance	*SE*	Percent	Variance	*SE*	Percent	Variance	*SE*	Percent
p	0.329	0.068	56.2	0.210	0.045	50.0	0.203	0.046	34.9
r	0.019	0.016	3.2	0.036	0.027	8.6	0.165	0.119	28.3
p x r, e	0.237	0.028	40.6	0.174	0.020	41.4	0.214	0.025	36.8

Note. $n = 73$.

Table 4

Estimated G-Coefficients and Φ-Coefficients for Average and Top 2 Scoring for each Task Based on the Number of Raters

Scoring Method	Number of Raters	Unusual Uses		Instances		Consequences	
		G	Φ	G	Φ	G	Φ
Average	1	0.70	0.63	0.73	0.64	0.54	0.34
	2	0.82	0.77	0.84	0.78	0.70	0.51
	3	0.87	0.83	0.89	0.84	0.78	0.61
	4	0.90	0.87	0.92	0.88	0.83	0.67
	5	0.92	0.89	0.93	0.90	0.86	0.72
Top 2	1	0.58	0.56	0.55	0.50	0.49	0.35
	2	0.74	0.72	0.71	0.67	0.66	0.52
	3	0.81	0.79	0.78	0.75	0.74	0.62
	4	0.85	0.84	0.83	0.80	0.79	0.68
	5	0.87	0.87	0.86	0.83	0.83	0.73

Table 5

Estimated Variance Components, Standard Error (SE)*, and Percentage of Variance Accounted for b*

Effects (Percent) for p x t for Uniqueness Scoring

Uniqueness Scoring (p x t)			
Effects	Variance	*SE*	Percent
p	0.724	0.272	15.9
t	1.294	0.939	28.5
p x t, e	2.529	0.292	55.6

Note. $n = 75$.

Table 6

G-Coefficients and Φ-Coefficients for Different Numbers of Tasks under Uniqueness Scoring, p x T

Number of Tasks	G	Φ
1	0.22	0.16
2	0.36	0.27
3	0.46	0.36
4	0.53	0.43
5	0.59	0.49
10	0.74	0.65
15	0.81	0.74
20	0.85	0.79

Table 7

Correlations Between Creativity Scores and Fluency Scores: Study 1

	r	*tau*
Average Creativity: Uses	-0.048	0.081
Top 2 Creativity: Uses	0.090	0.102
Uniqueness: Uses	0.474	0.381
Average Creativity: Instances	-0.228	-0.099
Top 2 Creativity: Instances	0.004	0.045
Uniqueness: Instances	0.672	0.505
Average Creativity: Consequences	-0.115	-0.058
Top 2 Creativity: Consequences	-0.184	-0.135
Uniqueness: Consequences	0.354	0.283

Note. The coefficients are Pearson *r* correlations and Kendall *tau* rank-order correlations. Note that *tau* coefficients tend to be lower than *r* coefficients for equivalent effect sizes (Gibbons, 1993).

Table 8

Descriptive Statistics and Correlations for Top 2 Scores: Study 2

	M	*SD*	1	2	3	4	5	6	7	8	9	10
1. Brick: Rater 1	2.14	.58	1									
2. Knife: Rater 1	2.12	.66	0.072	1								
3. Brick: Rater 2	1.12	.38	0.385	0.089	1							
4. Knife: Rater 2	1.17	.44	0.078	0.376	0.089	1						
5. Brick: Rater 3	2.41	.78	0.577	0.191	0.33	0.172	1					
6. Knife: Rater 3	2.54	.66	0.114	0.607	0.198	0.410	0.216	1				
7. N (NEO)	2.80	.65	0.012	0.046	0.059	0.037	-0.042	-0.026	1			
8. E (NEO)	3.56	.55	0.025	-0.016	-0.015	-0.013	0.024	0.000	-0.08	1		
9. O (NEO)	3.17	.53	0.131	0.103	0.116	0.055	0.084	0.084	0.074	0.013	1	
10. A (NEO)	3.75	.56	0.012	0.057	0.046	-0.037	0.051	0.028	-0.258	0.299	-0.007	1
11. C (NEO)	3.73	.59	-0.109	-0.119	0.051	-0.071	-0.133	-0.037	-0.274	0.19	-0.013	0.302
12. N (IPIP)	2.77	.85	-0.014	-0.032	0.048	0.024	-0.046	-0.102	0.749	-0.135	-0.016	-0.362
13. E (IPIP)	3.31	.87	0.133	-0.05	0.007	0.010	0.12	-0.005	-0.04	0.605	0.049	-0.101
14. O (IPIP)	3.43	.65	0.23	0.097	0.022	0.011	0.199	0.074	-0.096	0.148	0.486	-0.058
15. A (IPIP)	4.20	.55	0.051	0.025	0.096	-0.001	0.078	0.041	-0.022	0.412	0.139	0.589
16. C (IPIP)	3.49	.71	-0.126	-0.147	-0.037	-0.019	-0.127	-0.099	-0.205	0.151	-0.087	0.203
17. N (BBF)	2.34	.90	0.062	-0.024	0.044	0.056	0.003	-0.100	0.629	-0.087	0.03	-0.265
18. E (BBF)	3.40	1.08	0.193	-0.044	-0.048	-0.022	0.092	-0.018	-0.018	0.522	-0.088	-0.131
19. O (BBF)	3.97	.76	0.225	0.085	0.113	0.020	0.166	0.121	-0.104	0.353	0.421	0.065
20. A (BBF)	3.99	.76	0.111	0.022	0.094	-0.003	0.067	-0.026	-0.058	0.319	0.091	0.645
21. C (BBF)	4.15	.75	-0.123	-0.195	-0.017	-0.045	-0.17	-0.179	-0.148	0.135	-0.096	0.258
22. Arts Major	.093	.29	0.238	0.127	0.181	0.140	0.263	0.121	-0.007	0.012	0.167	-0.03

(table continues)

11	12	13	14	15	16	17	18	19	20	21	22
1											
-0.15	1										
-0.014	0.006	1									
0.083	-0.089	0.261	1								
0.302	-0.019	0.122	0.035	1							
0.785	-0.071	0.033	0.133	0.179	1						
-0.177	0.753	-0.009	-0.057	0.007	-0.083	1					
0.027	0.042	0.753	0.196	0.021	0.082	0.018	1				
0.114	-0.173	0.394	0.524	0.238	0.1	-0.103	0.295	1			
0.249	-0.167	-0.024	0.002	0.634	0.126	-0.139	-0.1	0.119	1		
0.733	-0.039	-0.007	-0.006	0.274	0.686	-0.158	0.017	0.058	0.228	1	
-0.149	0.013	0.005	0.166	-0.012	-0.123	0.065	0.008	0.181	0.023	-0.159	1

Note. $n = 226$. See text for abbreviations. Scores for all variables ranged from 1 to 5, except for Arts Major, which ranged from 0 to 1.

Table 9

A Summary of Explained Variance in Divergent Thinking: Study 2

Predictors	*Top 2 Scores*	*Average Scores*
NEOAC	49.4%	17.2%
Plasticity & Stability	42.1%	17.0%
NEOAC & Arts Major	57.6%	21.8%
Plasticity, Stability, & Arts Major	54.0%	22.0%

Note. NEOAC refer to the Big 5 domains; Plasticity and Stability refer to the higher-order factors of the Big 5. Percentages refer to the percentage of variance in divergent thinking explained by the predictors.

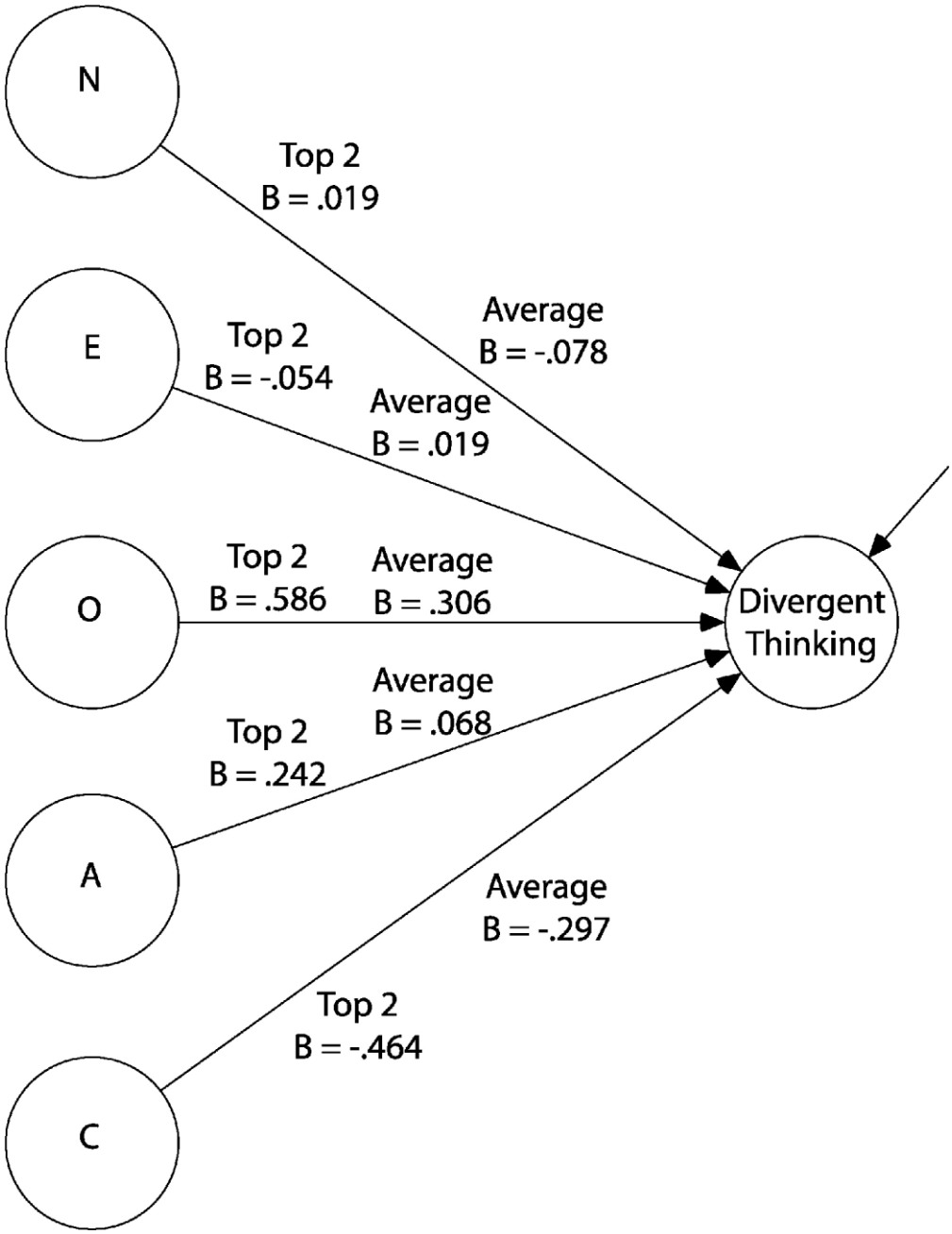

Figure 1. Predicting divergent thinking from the Big 5 factors.

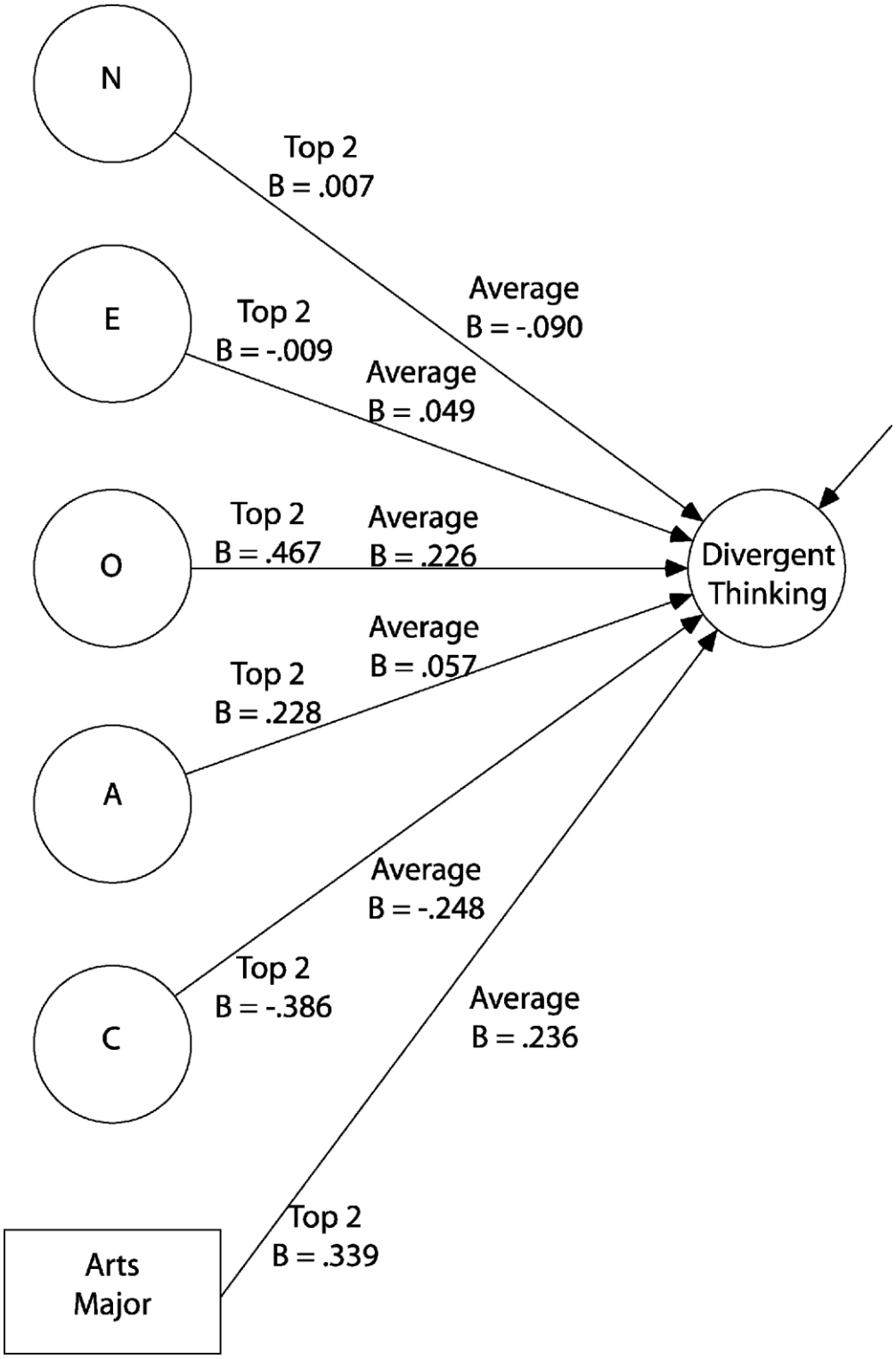

Figure 2. Predicting divergent thinking from the Big 5 factors and arts major.

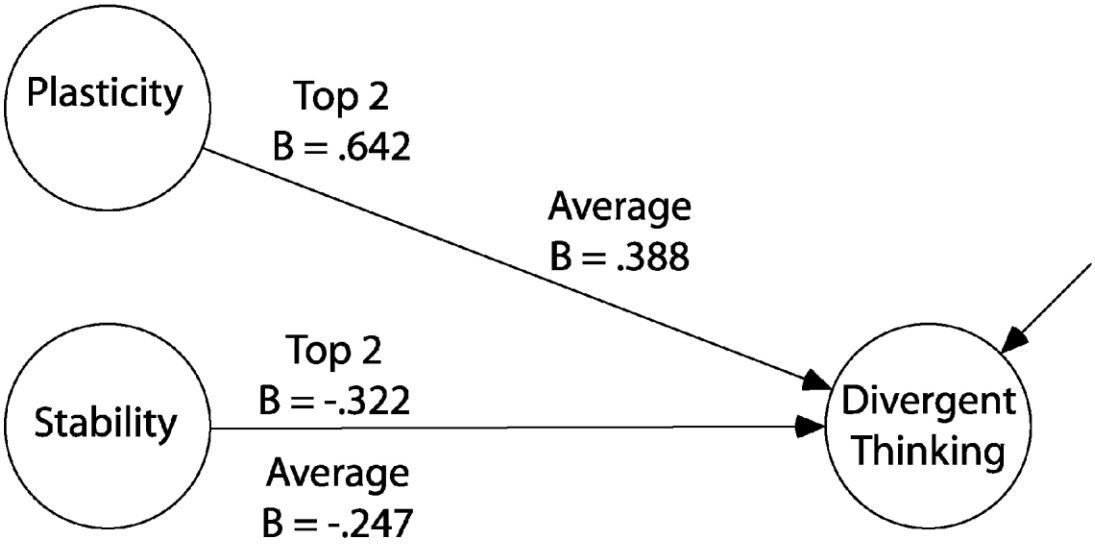

Figure 3. Predicting divergent thinking from the Huge 2 factors.

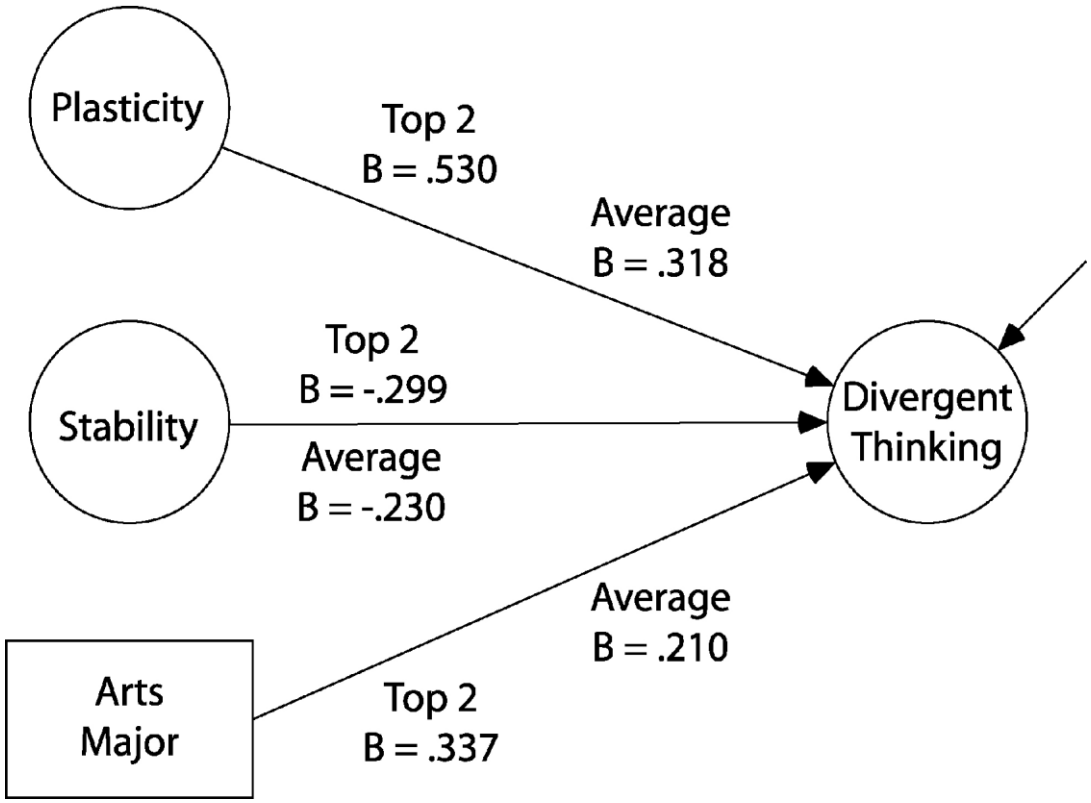

Figure 4. Predicting divergent thinking from the Huge 2 factors and arts major.

Appendix 1: Instructions for Judging Creativity

Creativity can be viewed as having three facets. Creative responses will generally be high on all three, although being low on one of them does not disqualify a response from getting a high rating. We will use a 1 to 5 scale:

1	2	3	4	5
not at all creative				*highly creative*

1. Uncommon

Creative ideas are uncommon: they will occur infrequently in our sample. Any response that is given by a lot of people is common, by definition. Unique responses will tend to be creative responses, although a response given only once needn't be judged as creative. For example, a random or inappropriate response would be uncommon but not creative.

2. Remote

Creative ideas are remotely linked to everyday objects and ideas. For example, creative uses for a brick are "far from" common, everyday, normal uses for a brick, and creative instances of things that are round are "far from" common round objects. Responses that stray from obvious ideas will tend to be creative, whereas responses close to obvious ideas will tend to be uncreative.

3. Clever

Creative ideas are often clever: they strike people as insightful, ironic, humorous, fitting, or smart. Responses that are clever will tend to be creative responses. Keep in mind that cleverness can compensate for the other facets. For example, a common use cleverly expressed could receive a high score.