Thinking Skills and Creativity 4 (2009) 79–85

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: http://www.elsevier.com/locate/tsc

A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking

Paul J. Silvia*, Christopher Martin, Emily C. Nusbaum

Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC 27402-6170, United States

ARTICLE INFO

Article history:
Received 18 March 2009
Accepted 26 June 2009
Available online 8 July 2009

Keywords:
Creativity
Divergent thinking
Validity
Psychological assessment
Openness to experience

ABSTRACT

Creativity assessment commonly uses open-ended divergent thinking tasks. The typical methods for scoring these tasks (uniqueness scoring and subjective ratings) are time-intensive, however, so it is impractical for researchers to include divergent thinking as an ancillary construct. The present research evaluated snapshot scoring of divergent thinking tasks, in which the set of responses receives a single holistic rating. We compared snapshot scoring to top-two scoring, a time-intensive, detailed scoring method. A sample of college students ($n = 226$) completed divergent thinking tasks and measures of personality and art expertise. Top-two scoring had larger effect sizes, but snapshot scoring performed well overall. Snapshot scoring thus appears promising as a quick and simple approach to assessing creativity.

Like parents, ombudsmen, and city council members, researchers are used to compromise. All assessment involves trade-offs between a method's evidence for validity and its cost. Many of the best assessment tools are costly in terms of administration time, expertise, technology, personnel-hours, and infrastructure. For this reason, many constructs have a range of available tools. A person's typical mood can be assessed with week-long experience-sampling methods or with brief self-report scales. Clinical symptoms can be assessed with face-to-face clinical interviews or with brief self-report screening scales. Even within a method, researchers can usually find a range of options. Personality researchers, for example, could choose the 300-item NEO-PI (Costa & McCrae, 1992), the 60-item FFI (Costa & McCrae, 1992), a 20-item IPIP scale (Donnellan, Oswald, Baird, & Lucas, 2006), or even one of two 10-item scales (Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2007).

The present research appraises a quick and simple method for assessing individual differences in creativity. Creativity research typically uses divergent thinking tasks to measure variation in creative abilities and potential (Kaufman, Plucker, & Baer, 2008; Plucker & Renzulli, 1999; Runco, 2007), but the traditional methods of coding and scoring these tasks are costly in terms of time and personnel. As a result, creativity is hard to include as a secondary or exploratory construct in a research project. We compare this brief, simple method—known as *snapshot scoring*—to a more time-consuming, detailed method (Silvia et al., 2008). If the brief method performs well, it may be a useful tool for researchers interested in assessing creativity.

1. Major approaches to scoring divergent thinking

Divergent thinking tasks assess creativity by asking people to generate ideas, which are then scored to capture variability in creativity (Plucker & Renzulli, 1999). For example, unusual uses tasks ask people to generate unusual uses for common objects,

* Corresponding author at: Department of Psychology, P.O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC 27402-6170, United States. Tel.: +1 336 256 0007; fax: +1 336 334 5066.
E-mail address: p_silvia@uncg.edu (P.J. Silvia).

1871-1871/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tsc.2009.06.005

such as bricks, knives, boxes, shoes, and paper clips. Over the decades, researchers have developed many methods for deriving scores from these tasks (Kaufman et al., 2008). Two scoring methods—uniqueness scoring and subjective ratings—have been used the most extensively.

1.1. Uniqueness scoring

Uniqueness scoring, formalized by Wallach and Kogan (1965) in their landmark research, is straightforward. Researchers compile all of the responses to a task and then assign each response a 0 or a 1. If a response is literally unique—if only one person in the sample gave the response—then it gets a 1. All other responses get 0s, regardless of how often they appeared. An appeal of this method is its implied definition of creativity: creative ideas are unique ideas. In this system, researchers focus on two scores: *uniqueness* (the number of unique responses) and *fluency* (the total number of responses), which indicate the quality and quantity of divergent thinking.

All research is at least slightly painful, but uniqueness scoring is uniquely painful. The responses must be transcribed into a lexicon, which may have tens of thousands of responses. Each response must then be compared with each other response to see if it appears only once or not. Along the way, many judgment calls must be made, such as whether "a brick walkway" and "a brick path" are unique responses and whether incomprehensible yet unique responses (e.g., "You know, that thing that hangs on a tree trunk. You know what I'm talking about, right?") deserve a point.

Uniqueness scoring has two major problems that have motivated the development of alternative assessment systems. The first problem is that uniqueness scores are highly correlated with fluency scores. Researchers pointed this out soon after Wallach and Kogan (1965) published their landmark tasks and scoring methods (e.g., Clark & Mirels, 1970), and later research further demonstrated that uniqueness and fluency are essentially confounded (Hocevar, 1979a, 1979b; Hocevar & Michael, 1979). This confounding is apparent even in the field's "gold standard" samples. In a reanalysis of Wallach and Kogan's data, Silvia (2008b) found a correlation of $\beta = .89$ between the latent fluency and uniqueness variables. In the large norm sample for the Torrance Tests of Creative Thinking (Torrance, 2008), fluency and uniqueness correlated equally highly (median $r = .88$). When two variables covary this strongly, it is hard to believe that they convey distinct information. Many contemporary researchers, in fact, use only the fluency scores as markers of creativity (e.g., Batey, Chamorro-Premuzic, & Furnham, 2009; Preckel, Holling, & Wiese, 2006).

The second problem is that uniqueness scores are strongly biased by sample size—as the sample becomes larger, creative responses become less common. Creativity thus becomes harder to detect, and the sample's estimated level of creativity declines, as the sample increases. In a small sample of 20 people, for example, most responses will be unique because the pool of responses is small. If 200 cases are added, however, many formerly unique responses will no longer be unique in the larger pool. In a huge sample, very few responses will be unique; it is theoretically possible, in fact, for no responses to be unique. Uniqueness scoring thus introduces a powerful sample dependence in its estimates of creativity, and this is bad for obvious reasons. Any assessment method that fares poorly in large samples is undesirable for basic research, unsuitable for large-scale testing programs, and unacceptable for high-stakes purposes (e.g., placement into gifted-education programs).

1.2. Subjective ratings of responses

A second assessment tradition uses judges to provide subjective ratings of creativity. The best-known example is the consensual assessment technique (CAT; Amabile, 1982), which is used to appraise creative products. In a typical CAT study, people are asked to generate poems, collages, or stories; experts in the area then judge the products on various dimensions (e.g., creativity, technical skill) according to their personal definition of creativity (Kaufman et al., 2008). Recent work has loosened the classic CAT to include diverse kinds of products, raters, and judgments (Baer, Kaufman, & Gentile, 2004; Kaufman, Gentile, & Baer, 2005; Kaufman, Lee, Baer, & Lee, 2007).

In divergent thinking research, it is difficult to conceive of what an "expert" divergent thinking judge would be. Instead of noted experts, stoic research assistants are usually pressed into service for the tedious process of rating divergent thinking responses. The responses are usually rated on a creative-quality scale, such as a 5-point "not at all creative" to "very creative" response format (Silvia et al., 2008). Unlike the CAT, which avoids training the raters, divergent thinking research usually provides some training, guidelines, or rubrics for judges, thus increasing between-judge agreement.

Subjective scoring of divergent thinking tasks dates at least to Guilford's seminal research; many of his studies used subjective ratings of originality, remoteness, and cleverness (e.g., Christensen, Guilford, & Wilson, 1957; Wilson, Guilford, & Christensen, 1953). Since then, many studies have used some kind of subjective quality rating. The most common approach is to have raters give a score to each response (e.g., Gilhooly, Fioratou, Anthony, & Wynn, 2007; Harrington, 1975; Silvia & Phillips, 2004). These scores can then be averaged for an overall score for the task.

When scoring each response, researchers can score other features of the responses and test if those features covary with the creativity ratings. For example, people's later responses are usually more creative than their first few responses (Christensen et al., 1957), responses that people generated on the spot are more creative than responses retrieved from memory (Gilhooly et al., 2007), and the two responses that people picked as their "top two" are more creative than the rest (Silvia, 2008c). Responses can also be coded for their length, elaborateness, concreteness, cleverness, and their clustering into classes—these open-ended tasks provide a lot of information.