Thinking Skills and Creativity 23 (2017) 129–139

Contents lists available at ScienceDirect

Thinking Skills and Creativity

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

Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores

Boris Forthmann[a,*], Heinz Holling[a], Nima Zandi[a], Anne Gerwig[a], Pınar Çelik[b], Martin Storme[b], Todd Lubart[b]

[a] *Westfälische Wilhelms-Universität Münster, Germany*
[b] *Université Paris Descartes, France*

ARTICLE INFO

Article history:
Received 16 May 2016
Received in revised form 15 November 2016
Accepted 14 December 2016
Available online 21 December 2016

Keywords:
Creativity
Divergent thinking
Ideational pools
Snapshot scoring
Rater cognition
Cognitive workload

ABSTRACT

Using a rater cognition approach, three extant datasets from recent divergent thinking research were used to analyze the use of subjective processes to rate the quality of ideas. Subjective ratings have gained popularity recently and often three classic dimensions are combined into a single score: uncommonness, remoteness, and cleverness. Thus, scoring of ideas or sets of ideas is a demanding task, in particular when a set contains many ideas. In such a situation, cognitive load is expected to be highest and errors are more likely. Using a cumulative ordinal logit model, results suggest that rater disagreement is predicted by the amount of information (complexity) that was coded. Rater disagreement was higher when participants were instructed to be creative (vs. standard instruction) and also a significant interaction of complexity and instruction was found. Simple slope analysis indicated that the influence of complexity on disagreement was less pronounced with a be-creative instruction and that the difference in disagreement between instructions was more pronounced for low-complexity as compared to high-complexity idea sets. Several implications for deriving subjective creativity ratings and training raters are discussed.

The assessment of creativity has been a subject of debate since years (Hocevar, 1981; Michael & Wright, 1989; Mouchiroud & Lubart, 2001; Silvia et al., 2008). The ongoing discussion focuses often on the creativity of products. Researchers have often relied on subjective ratings (Amabile, 1982; Chen et al., 2002; Kornilov, Kornilova, & Grigorenko, 2016; Sternberg, 2012) and more recently the behavior of raters when judging creativity has been the focus of several studies (Long, 2014; Long & Pang, 2015; Storme, Myszkowski, Çelik, & Lubart, 2014; Tan et al., 2015). For example, Long (2014) examined the criteria that were used by judges to rate creativity in science tasks and Long and Pang (2015) examined rater cognition when the same science tasks were rated.

However, the creativity of responses to convergent or divergent thinking tasks has also been considered in creativity research. In relation to this, it is widely believed that the quantity of creative performance, such as the number of ideas in a divergent-thinking task, does not tell the full story. A measure of quality of ideas is considered to add valuable information in research or diagnostic practice. Such quality measures have been taken into account right from the start of creativity

* Corresponding author at: Institut für Psychologie, Westfälische Wilhelms-Universität, Fliednerstrasse 21, 48149 Münster, Germany.
E-mail address: boris.forthmann@wwu.de (B. Forthmann).

http://dx.doi.org/10.1016/j.tsc.2016.12.005

measurement research (Guilford, 1956) and subjective scoring methods (Mouchiroud & Lubart, 2001; Runco & Mraz, 1992), which are accepted to be valid for products (Amabile, 1982; Baer, 2008), were more recently again discussed and analyzed in detail for divergent-thinking tasks (Silvia, 2011; Silvia et al., 2008; Silvia, Martin, & Nusbaum, 2009). Following this work, these methods were used by other researchers in several studies on different issues related to divergent thinking (for example, Hass, 2015; Hofelich Mohr, Sell, & Lindsay, 2016; Lee & Therriault, 2013).

Generally, subjective scoring procedures of divergent-thinking tasks are not applied in the same way and most often the scoring dimensions vary from study to study. For example, idea response sets or ideas of divergent-thinking tasks were judged for overall creativity (Diedrich, Benedek, Jauk, & Neubauer, 2015; Mouchiroud & Lubart, 2001; Runco & Charles, 1993; Runco & Mraz, 1992; Storm & Patel, 2014), originality or unusualness (Mouchiroud & Lubart, 2001; Runco & Charles, 1993; Storm & Patel, 2014), novelty (Diedrich et al., 2015; Olteteanu & Falomir, 2016; Storm & Patel, 2014), cleverness (French, Ekstrom, & Price, 1963; Mullins, 1963), and usefulness or appropriateness (Diedrich et al., 2015; Olteteanu & Falomir, 2016; Runco & Charles, 1993).

Moreover, the snapshot scoring approach proposed by Silvia et al. (2009) involves the integration of three dimensions, *uncommonness*, *remoteness*, and *cleverness*, to derive a score of creative quality for a set of a person's ideas. These dimensions date back to the classic work of Wilson, Guilford, and Christensen (1953) on indicators of originality; such indicators were interpreted by Guilford (1966) as signs of creative quality and studies in his laboratory found these dimensions to load on the same factor (for example, Kettner, Guilford, & Christensen, 1959; Wilson, Guilford, Christensen, & Lewis, 1954). Moreover, combinations of these dimensions were then used in later studies for subjective quality scores in divergent thinking (for example, Hocevar, 1979; Zarnegar, Hocevar, & Michael, 1988).

We used snapshot scoring in our own research (Forthmann, Gerwig, et al., 2016) and observed during extensive training sessions that cognitive workload can lead to errors in the scoring process. In other words, some ideas of a set can be overlooked, especially when a set contains many ideas. Consequently, the overall reduced workload for the raters (one score for the whole set vs. as many scores as ideas are in a set) does not necessarily mean that deriving a score for an idea set is equally demanding as deriving a score for a single idea. It can be argued that the workload to integrate all three dimensions over a rather large set of ideas, which is perhaps also heterogeneous regarding the creativity of single ideas, can be even greater than rating all of these ideas separately. Moreover, even when single ideas are rated, differences in cognitive workload are expected because ideas are expressed with varying word lengths. Therefore, the link between rater disagreement and cognitive workload might even show up when single ideas are rated.

Thus, analyzing the influence of the amount of information that needs to be processed by raters to derive a creativity score and the possible effect on rater disagreement will be examined in the current article.

1. Possible effects of divergent thinking instructions on rater performance

Instruction manipulations have a long tradition in divergent thinking research (Harrington, 1975; Runco, Illies, & Reiter-Ralmon, 2005; Runco & Okuda, 1991) and have been shown to influence also other creativity tasks (Chen et al., 2005). More recently, Nusbaum, Silvia, and Beaty (2014) discussed possible differences of what is being measured depending on the kind of instruction that is used. They suggest that instuctions which set the focus on production (as many ideas as possible) make divergent-thinking tasks more similar to verbal fluency tasks (Silvia, Beaty, & Nusbaum, 2013; Unsworth, Spillers, & Brewer, 2010). Consequently, instructing participants to be creative is considered to be a methodological necessity in order to validly measure divergent thinking. However, very often a standard as-many-as-possible instruction (Runco & Acar, 2010; Runco et al., 2005) or a hybrid instruction (Madore, Jing, & Schacter, 2016) is used for divergent thinking research. A typical pattern of results depending on the given instruction is a decline of fluency and an increase of quality with a be-creative instruction as compared to an instruction with a production focus (Nusbaum et al., 2014). Thus, analyzing the effect of instruction on rater performance is important for two reasons: (a) Instructions with a production focus are most often used and be-creative instructions are considered as a methodological requirement, thus both instructions have importance for the field and (b) instructions affect the nature of the given response sets in such a way that the amount and kind of information in the sets can be expected to be different.

Consequently, instructions should be taken into account when rater effects are analyzed. Is it easier for the raters to grasp the creativity of an idea set if there are less ideas (less information) in a set generated with a be-creative instruction? Or are those fewer ideas particularly complex in this case and, as a consequence, ratings are harder? In addition, it could be the case that the ratings with a production instruction are easier because the distribution of high-quality ideas in divergent thinking is generally less scattered (Harrington, 1975; Runco, 1986) due to the participants unconstrained idea generation and, thus, the non-original content is easier to encode. In addition, instruction may not only have a main effect on rater-disagreement. The type of instruction may play a moderating role for the relationship of amount of information and rater-disagreement. For example, more ideas are expected with a many-as-possible instruction (for example, Nusbaum et al., 2014; Runco, 1986) and due to the unconstrained idea generating process there is no need to waste time with complex expressions and, therefore, more information here could be a reflection of many simple ideas. However, following a be-creative instruction, more information in an idea or idea set is likely to be due to more details. In line with this argument