Published in final edited form as:
Psychol Aesthet Creat Arts. 2013 November 1; 7(4): 341–349. doi:10.1037/a0033644.

Assessment of Divergent Thinking by means of the Subjective Top-Scoring Method: Effects of the Number of Top-Ideas and Time-on-Task on Reliability and Validity

Mathias Benedek, **Caterina Mühlmann**, **Emanuel Jauk**, and **Aljoscha C. Neubauer**
Department of Psychology, University of Graz, Austria

Abstract

Divergent thinking tasks are commonly used as indicators of creative potential, but traditional scoring methods of ideational originality face persistent problems such as low reliability and lack of convergent and discriminant validity. Silvia et al. (2008) have proposed a subjective top-2 scoring method, where participants are asked to select their two most creative ideas, which then are evaluated for creativity. This method was found to avoid problems with discriminant validity, and to outperform other scoring methods in terms of convergent validity. These findings motivate a more general, systematic analysis of the subjective top-scoring method. Therefore, this study examined how reliability and validity of the originality and fluency scores depend on the number of top-ideas and on time-on-task. The findings confirm that subjective top-scoring avoids the confounding of originality with fluency. The originality score showed good internal consistency, and evidence of reliability was found to increase as a function of the number of top-ideas and of time-on-task. Convergent validity evidence, however, was highest for a time-on-task of about 2 to 3 minutes and when using a medium number of about three top-ideas. Reasons for these findings are discussed together possible limitations of this study and future directions. The article also presents some general recommendations for the assessment of divergent thinking with the subjective top-scoring method.

Keywords

Originality; fluency; time-on-task; reliability; validity

Divergent thinking tests have enjoyed long-standing popularity in creativity research but also have faced persistent debates about their limitations. A common issue is that divergent thinking ability may be considered a useful indicator of creative potential, but it may not generalize to a more general conceptualization of creativity which e.g. also includes real-life creative achievement (Runco & Acar, 2012). A second common issue is related to the unsatisfactory psychometric properties (i.e., objectivity, reliability, and validity) of divergent thinking scores. These psychometric issues need to be resolved in order to establish confidence in the use of divergent thinking tasks for the assessment of creative potential and for the study of the cognitive and neurocognitive mechanisms underlying creative ideation

Correspondence concerning this article should be addressed to Mathias Benedek, Department of Psychology, University of Graz, Austria. Postal address: Maiffredygasse 12b, 8010 Graz, Austria. mathias.benedek@uni-graz.at.

Europe PMC Funders Author Manuscripts

(e.g., Benedek, Könen & Neubauer, 2012; Fink & Benedek, in press; Gilhooly et al., 2007; Nusbaum & Silvia, 2012). In the past few years, strong efforts have been made to further examine divergent thinking tests in the light of different methodological considerations and to propose new solutions to common issues (e.g., Plucker, Qian & Wang, 2011; Runco, Okuda and Thurston, 1987; Silvia, Martin & Nusbaum, 2009; Silvia et al., 2008). This study aims to extent these developments by a systematic examination of the psychometric properties of subjective scoring methods.

Divergent thinking tasks require participants to generate creative solutions to given open problems. A large number of different divergent thinking tasks have been devised (e.g., the alternate uses tasks ask to find creative uses for a commodity item such as a brick; cf., Benedek, Fink & Neubauer, 2006), and a variety of different measures have been proposed for scoring responses generated in these tasks (e.g., Torrance, 2008). These measures commonly involve a scoring of the fluency and originality of ideas, which can be considered to reflect the quantity and quality of ideation performance. The scoring of ideational fluency is straightforward as it essentially requires counting the number of relevant responses. In contrast, the scoring of originality is more complex and can be achieved by different methods. In the uniqueness scoring, the originality of responses is defined by their statistical infrequency. For example, infrequent responses ($p < 5$ - 10%) are usually defined as unusual or unique, whereas more frequent responses are considered to be common (e.g., Runco, 2008; Torrance, 1974). The originality score then is obtained by counting the number of unique responses. While this method appears to allow for an objective scoring, a number of serious objections have been raised including the issue that statistical infrequency may not be a valid indicator of creativity since it does not account for the appropriateness of responses (Silvia et al., 2008). As an alternative, in the subjective scoring method, external judges are employed to evaluate all responses for creativity (i.e., unusualness and appropriateness; cf., Amabile, 1982) and ratings are finally summed. Good interrater reliability of this method can be seen as an argument for a certain objectivity of this method, but the evaluation of large amounts of responses by different judges is still very laborious.

The uniqueness scoring and the subjective scoring method, however, also face a more general methodological issue. Ideational fluency has been realized to act as a contaminating factor for all other scores (Hocevar, 1979a, 1979b; Kaufman, Plucker & Baer, 2008; Michael & Wright, 1989; Runco et al., 1987). According to the scoring techniques outlined above, the scoring of originality is directly related to the number of responses (i.e., fluency score). A person who gives more responses thus is more likely to get points for originality. This explains for the extremely high correlations of fluency with originality scores, which often range from $r = .80$ to .90 (e.g., Mouchiroud & Lubart, 2001; Torrance, 2008). It has been argued that these marked correlations do not support discriminant validity (Plucker et al., 2011; Silvia et al., 2008). Moreover, after the effect of fluency is partialled out, the reliability evidence of the originality score is usually very low (Hocevar, 1979a, 1979b; Runco et al., 1987); one study found that reliability is still adequate for gifted children performing figural tasks (Runco & Albert, 1985). The reliability and validity of originality scores hence appear to be substantially affected by the correlation with ideational fluency.

Since ideational originality is conceived as an essential qualitative factor of divergent thinking ability, a number of suggestions were made on how to control for the confounding influence of ideational fluency. One suggestion is that the evaluations should be based on the entire set of ideas rather than single ideas (i.e., scoring of ideational pools, or snapshot scoring; Runco & Mraz, 1992; Silvia et al., 2009). This method allows for a very quick overall assessment but was found to yield only moderate evidence of reliability. It was also proposed to divide total originality by the number of ideas (i.e., average scoring, or ratio scoring). This method has some merits (e.g., Plucker et al., 2012; Silvia et al., 2008), but again it sometimes was found to show very low reliability evidence (Runco, Okudo & Thurston, 1987), and it should be noted that average originality might not be valid for the ability to come up with the most creative ideas. Another possibility is to focus on a constant number of responses (Clark & Mirels, 1970). The examinees can, for example, be instructed to produce a predefined number of responses (e.g., generate three creative responses; Hocevar, 1979b). This method controls for fluency, but no longer allows for the implicit assessment of fluency. As an alternative, the scoring can be restricted to a predefined number of responses from the entire response set (Michael & Wright, 1989). This method can be called *subjective top-scoring*. Recently, Silvia et al. (2008) have adopted this approach by proposing the top-2 scoring method. This method asks the examinees to indicate their two most creative responses per task, and only these two responses then are evaluated. The top-2 scoring of originality was shown to avoid excessive correlations with fluency and to perform better than the snapshot scoring or average scoring (Silvia, 2011; Silvia et al., 2008, 2009). Similarly, Reiter-Palmon, Illies, Cross, Buboltz and Nimps (2009), using more complex, real-life divergent thinking tasks, found that using the single most creative response (i.e., top-1 scoring) is suitable to overcome a confounding with fluency.

Evaluating people by their best responses reflects a maximum performance condition (Runco, 1986) and acknowledges that the ability to select one's best ideas is important for creativity (Smith, Ward & Finke, 1995). This may involve that generative and evaluative processes become confounded (Runco, 2008), but examinees were found to be quite discerning in selecting their most creative ideas which supports the validity of this procedure (Silvia, 2008). Finally, from a practical point of view, this method also enhances the efficiency of the rating procedure. Silvia et al. (2008) reported that by using top-2 scoring only about 28% of the total number of ideas had to be evaluated.

Recently, Plucker et al. (2011) have compared different methods of originality scoring with respect to reliability and validity. The methods included uniqueness scoring, average uniqueness (i.e., dividing uniqueness by fluency), uniqueness of the first or last 10 ideas and subjective, rater-based scorings of the entire response set (summative score) or the first or last 10 ideas. Using only two items of the instances task, reliabilities ranged between .37 and .62. The average uniqueness score was found to show somewhat higher correlations with self-report creativity measures and negative correlations with fluency. It was concluded that this method could be favored over subjective rater-based methods. However, since in this study participants were not asked to select their most creative ideas, the authors suggested that examining the reliability and validity of top-ideas "is also promising and should be the subject of additional study" (p. 15).

Europe PMC Funders Author Manuscripts

Main Research Questions

The main idea underlying subjective top-scoring of originality is to focus on a constant number of top-ideas in order to avoid a confounding with fluency which questions its discriminant validity. The top-2 scoring method, which focusses on the two most creative ideas, was found to perform well in terms of reliability and validity as compared to other scoring methods such as uniqueness scoring (Silvia et al. 2008). In recent investigations, we have also employed a subjective scoring of divergent thinking tasks and found that correlations with intelligence crucially depended on the scoring method (i.e., top-2 vs. average originality; Jauk, Benedek & Neubauer, under review). Moreover, we observed that top-3 scoring resulted in a somewhat higher reliability evidence of the originality score as compared to top-1 or top-2 scoring (Benedek, Franz, Heene & Neubauer, 2012). This raises the question to what extent the number of top-ideas actually affects reliability and maybe also the validity of the originality score. Moreover, since the number and originality of ideas depend on time-on-task (e.g., Beaty & Silvia, 2012; Mednick, 1962), the most adequate number of top-ideas might also depend on the duration of divergent thinking tasks. Therefore, this study aims to have a close look on different realizations of the subjective top-scoring method and their effects on the psychometric properties of originality scores. Specifically, we want to examine systematically to what extent a) the actual number of top-ideas and b) the time-on-task affect 1) the correlation of originality scores with fluency, 2) the reliability of originality scores, and 3) the convergent validity of originality scores. Additionally, we also examine the effect of task duration on the reliability and validity of fluency scores. We thereby hope to reveal further information about the adequate assessment of ideational originality, ensuring high psychometric quality but also efficient scoring procedures.

Method

Participants

A sample of 105 participants (51 females) took part in this study. The age ranged from 18 to 51 years (M = 23.80, SD = 3.97). 49% of participants were students of Psychology at the University of Graz, 38% were majoring in different fields, and the remaining 13% were non-students. Participants were invited to take part in a study on creativity and personality and were offered credits for participation in empirical investigations (if applicable) and an individual feedback on personality structure in exchange for participation. The only requirement for participation was basic computer literacy. All participants gave written informed consent. The study was approved by the local ethics committee.

Tasks and Material

We employed six divergent thinking tasks timed for five minutes each. The tasks included three alternate uses tasks (“car tire”, “glass bottle”, and “knife”) and three instances tasks (“what could be round?”, “what could make a loud noise?”, “what could be used for faster locomotion”). Tasks were administered by a self-devised computer program written in Matlab (The Mathworks; Natick, MA), which allows for acquisition of time-stamped

responses. There is evidence that computer-based assessment of divergent thinking is highly comparable to a paper-pencil assessment (Lau & Cheung, 2010).

In an initial general instruction participants were told that they will be presented some questions for which they should try to "generate as many different unusual and creative responses as possible". They were asked to express their ideas as succinctly as possible, and to write each idea into the input box and then press the enter-key to add it to their idea list. Participants were told that there was "some minutes" time for each task and that the program would proceed automatically as soon as time is over. By giving participants no exact information about the total or remaining task duration, we hoped that they would keep on entering every idea as soon as it comes to mind, but not to develop specific task strategies related to a five minutes task time. After a task was started, participants were presented the specific task instructions on top of the screen (e.g., "What could make a loud noise? Name all the unusual and creative responses that you can think of."). Below, there was an editable input box where ideas could be entered. Every idea was added to a list placed below the input box. Two time events were recorded for each idea: 1) the time when the participant started entering the idea, and 2) the time when writing was complete and the idea was added to the list. We only considered the former time event, since this can be considered as the time when the idea actually came to mind, whereas the latter time event depends on the length of the idea and the typing speed.

After all tasks were completed, ideas were ranked for creativity by the participants. To this end, participants were presented with lists showing all their ideas within a single task with the ideas being arranged in randomized order. They were asked to rearrange the position of the ideas until the sequence of ideas in the list reflected the creativity of ideas as subjectively appraised by the participants. Ideas were rearranged by selecting them and moving them up or down by means of specific buttons. At the end, the topmost idea in the final list should be the most creative one, the second idea in the list should be the second-most creative one, and the last idea in the list should be the least creative one. This was done for all six tasks, separately.

We also measured self-reported ideational behavior by means of a German version of the Runco Ideational Behavior Scale (RIBS; Runco, Plucker & Lim, 2000). Personality structure was assessed by means of the five-factor inventory NEO-FFI (Borkenau & Ostendorf, 1993).

Scoring of Divergent Thinking Tasks

The ideas generated in the divergent thinking tasks were scored for fluency and originality. Fluency scores simply reflect the number of ideas generated after a given time. Originality scores were computed according to the subjective top-scoring method using the creativity evaluations obtained by external judges.

External originality ratings—Participants generated a total of 10921 ideas in the six divergent thinking tasks. All ideas were pooled and identical ideas were removed resulting in a final set of non-redundant 6229 ideas. Eight external raters were asked to evaluate the creativity of the ideas on a scale ranging from 0 ("not creative") to 3 ("very creative"). All

raters received an initial training, which made them familiar with the scale (e.g., they were informed that ideas can be considered “highly creative” when they are perceived as original and useful, and probably only few people will come up with them). The judges evaluated a small subset of ideas and after that discussed their ratings. Due to the large amount of ideas, each judge then evaluated the ideas of only half of the tasks so that finally there were four independent ratings for each idea. The interrater reliability between the four judges was ICC = .68, .80, .65, .60, .51, and .68 for the tasks “car tire”, ‘glass bottle”, ‘knife”, “round”, “loud noise”, and “faster locomotion”, respectively. The creativity of a single idea was defined as the average creativity rating given by the four external judges.

Subjective top-scoring—The main idea of the top-scoring method is that the originality score is based on the creativity evaluations of a predefined number of top-ideas. The top-ideas are identified by the participants themselves according to their subjective appraisal of the creativity of their ideas. For example, Silvia et al. (2008) employed a top-2 scoring, where participants marked their two most creative ideas which then were considered for scoring. Generally, all kinds of top-scores can be computed. For example, for a top-1 score only the single most creative idea within a task would be considered, whereas for a top-3 score the three most creative ideas would be included.

A first specific aim of this study was to examine the effect of the number of *top-ideas*. This was made possible by having participants sort all their ideas for creativity, as that allows for a post-hoc classification of any number of top-ideas. Second, we aimed to also consider the effect of *time-on-task*. In this study, time-on-task can theoretically vary from zero to five minutes (i.e., the total time for each task). For a specific time-on-task lower than five minutes, e.g. 3 minutes, the scores were computed based on the data available after 3 minutes. To illustrate this method, let us consider the following example: Assume that a participant generated four ideas at 30, 60, 120 and 240 seconds, and afterwards ranked them 2., 4., 3., 1. (i.e., the first idea being second-most creative, the second idea being least creative, and so on). Then, for computing the *top-2* originality score for a *time-on-task* of 3 minutes, only the two most creative ideas within the first 3 minutes are considered, hence, the first and the third idea. The originality score was finally computed by averaging the creativity evaluations of the considered top-ideas. If a participant generated fewer ideas than the number of top-ideas then the creativity evaluations of the available ideas were averaged.

Procedure

Participants were tested in small groups of up to five people in a computer room. They first performed the six divergent thinking tasks, which were presented in a randomized order. The divergent thinking tasks were preceded by a short exercise (enter words starting with the letter “F”) to become familiar with the general procedure of a computer-based idea generation task. After completion of the six tasks, the participants ranked their ideas for creativity. Finally, the participants completed the personality inventory, and the ideational behavior scale. The whole session took about one hour.

Europe PMC Funders Author Manuscripts

Analysis Plan

The main analyses include correlation analysis of fluency and originality scores, reliability analyses (i.e., internal consistency of scores), and validity analyses (i.e., correlations with external criteria). In all of these analyses two experimental factors are varied systematically: First, the factor *top-ideas* is varied from 1 to 10 ideas (see section on top-scoring method for further details). Additionally, this factor also includes the value “all ideas”, where all ideas given by a participant were considered; this hence corresponds to an average originality score (cf., Silvia et al. 2008). This factor only applies for the originality score but not for the fluency score. Second, the factor *time-on-task* is varied from 1 to 5 minutes (i.e., scores are computed for time-on-task of 1, 2, 3, 4, and 5 minutes). In total, the scoring hence was computed for 55 different conditions (11 top-idea conditions by 5 time-on-task conditions). The results of these analyses are visualized by means of contour plots (see Figures 1, 2, and 3).

Results

Descriptive Statistics and Preliminary Analyses

Participants generated on average 17 ideas (SD = 6.36) within the task time of 5 minutes. The fluency of ideas was significantly higher in the three instances tasks (M = 21.95, SD = 8.61) than in the three alternate uses tasks (M = 12.72, SD = 4.96), $t(104) = 15.83$, $p < .01$. As can be seen in Table 1, the total number of ideas steadily increases over time, however, the fluency of ideas also steadily declines over time starting from the first minute. Considering the alternate uses tasks, half of participants had created four ideas or more within the first minute of the task but the increase flattened down to one additional idea in the last minute of the task (difference between median values after 4 and 5 minutes; see Table 1). It can also be seen that there are large individual differences in fluency scores. While the least fluent 10% of participants generated only 9 ideas or less after working on the instances task for 5 minutes, the most productive 10% of the sample generated more than three times as many ideas (i.e., 32.7).

A principal factor analysis with Varimax rotation and Kaiser normalization was performed for the six fluency scores and the six average originality scores derived from the alternate uses as well as the instances tasks. The analysis extracted two factors (according to the Kaiser criterion and according to the Scree test; K-M-O = .83), explaining 67% of total variance. Further evidence for a two-factorial solution comes from the minimum average partial test (MAP test: Velicer, Eaton, & Fava, 2000), which returned two components as the number of factors to extract from the twelve measures. This two-factor solution clearly revealed a fluency factor and an originality factor with all six fluency and originality scores loading on corresponding factors (unspecific loadings were below .25). In other words, we obtain evidence for score-specific factors rather than task-specific factors, which supports the feasibility of aggregating fluency and originality scores across tasks.

How Do Scoring Conditions Affect the Correlation of Fluency and Originality?

We computed correlations between originality and fluency scores for different numbers of top-responses at varying time-on-task. It was assumed that the subjective top-scoring

method can avoid excessively high correlations between originality and fluency since it focusses on a constant number of ideas. In line with these expectations, correlations were found to be close to zero ranging between −.30 and .06 (see Figure 1). Ideational fluency and originality even showed significant negative correlations when using the average score and considering task times of 3 minutes or less.

For reasons of comparison with previous studies, we also computed summative originality scores by summing up the creativity evaluations of all ideas produced by participants within a task. As expected, these summative creativity scores showed extremely high correlations with the fluency scores of r = .83, .87, .90, .90, .91 for time-on-task of 1, 2, 3, 4, and 5 minutes, respectively.

How Do Scoring Conditions Affect Reliability?

To examine how the scoring conditions affect reliability of fluency and originality scores, we computed their internal consistency (Cronbach's α). The fluency score shows high reliability (α = .83) even for a short time-on-task of 1 minute. The reliability evidence increases with time-on-task up to α = .89 for a time-on-task of 5 minutes (see Figure 2, a).

The reliability evidence of the originality score also was found to generally increase with an increasing number of top-ideas and with increasing time-on-task (see Figure 2, b). Reliability was lowest when only a single top-idea was considered (i.e., top-1 score) staying below an alpha of .60, but it increased substantially by including some additional top-ideas. For example, at a time-on-task of 2 minutes using top-2, top-3, or top-4 scoring increased reliability to .70, .75, or .77, respectively. Time-on-task also generally increased reliability evidence but this is especially true when a larger number of top-ideas is considered. For the top-1 scoring, the increase of time-on-task from 1 minute to 5 minutes only causes an increase in reliability from .56 to .59, whereas it increases from .71 to .83 for the top-5 scoring. A decent alpha coefficient of at least .75 could be obtained only by using a time-on-task of 2 minutes (or higher) and when using at least the top-3 ideas. An alpha of .80 was obtained when using top-4 scoring with a time-on-task of 4 minutes or top-5 scoring with time-on task of 3 minutes. Reliability of the originality score peaked at an alpha of .87 when using the average score (i.e., using all ideas) at a time-on-task of 5 minutes.

How Do Scoring Conditions Affect Validity?

The effect of the scoring method on the convergent validity evidence of ideational fluency and originality was tested by means of correlations with the external criteria of self-reported ideational behavior and the personality factor openness. The fluency score showed significant positive correlations with both external criteria ranging from .26 to .33 for the ideational behavior scale and from .25 to .30 for openness, respectively (see Figure 3, a and c). At this, there is a small trend for correlations to increase with time-on-task.

The originality score showed no significant correlations with the ideational behavior scale, but just a weak trend towards positive correlations (see Figure 3, b). With respect to openness, the originality scores generally showed significant positive correlations (see Figure 3, d). These correlations were highest (r = .35 to .38) for a time-on-task of 2 minutes

Europe PMC Funders Author Manuscripts

when using scoring of top-2 to top-8. The correlations were substantially lower but still significant (r = .21 to .26) for the average scoring method which considers all ideas.

Discussion

Can Subjective Top-scoring Avoid the Confounding of Originality and Fluency?

One major aim of the subjective top-scoring method is to avoid the usually high dependency of qualitative measures of divergent thinking (e.g., ideational originality) from the number of ideas generated by participants (i.e., ideational fluency). We were able to replicate the common finding that a summative scoring of originality (i.e., computing a sum of the creativity evaluations of all ideas generated by a participant) results in extremely high correlation with the fluency scoring ranging between .80 and .90 (cf., Mouchiroud & Lubart, 2001; Torrance, 2008). In contrast, when originality scores were computed by means of the top-scoring method, correlations with fluency were largely close to zero. This is in line with the finding of Silvia et al. (2008), who also obtained no significant correlation with fluency when using top-2 scoring. The results hence confirm that the subjective top-scoring method avoids the confounding of originality scores with fluency.

The average score is a special case as it uses all ideas but by averaging ratings rather than summing them, a high positive correlation with fluency of ideas can be avoided. For the average score (and to a smaller extent also for the top-9 or top-10 score) we even observed small negative correlations at least when time-on-task was short. This result may probably be attributed to the existence of people who focus on fluency rather than creativity of ideas and thus were able to generate large amounts of responses. This strategy probably involves the generation of a large number of highly common responses which then results in a low average originality score as compared to those who rather focus on creativity of ideas (Reiter-Palmon et al., 2009).

Psychometric Properties of the Fluency Score

We obtained high internal consistency for the ideational fluency scores of the six divergent thinking tasks. Alpha coefficients slightly increased with time-on-task but already settled above .85 for times-on-task of 2 minutes or more. This suggests that ideational fluency can be reliably assessed even with short divergent thinking tasks. We further obtained significant positive correlations of fluency with self-reported ideational behavior and openness supporting the general validity of this score. These correlations also showed a slight increase with time-on-task which can probably be attributed to the corresponding increases in reliability.

Psychometric Properties of the Originality Score

The top-scoring method was found to result in dependable originality scores. Although interrater reliability was moderate for some tasks, the internal consistency between the six different divergent thinking tasks reached Cronbach’s alpha levels well beyond .80 for some scoring conditions. This level of reliability bears comparison with other well-established constructs of cognitive ability. Together with the findings derived from factor analysis, this indicates that originality scores coming from different divergent thinking tasks share a

substantial amount of common variance. Although divergent thinking tasks may not be fully interchangeable with respect to their cognitive demands (Guilford, 1967; Kuhn & Holling, 2009; Silvia, 2011), our results support the feasibility of computing aggregate scores across different divergent thinking task in order to obtain a reliable total originality score. The reliability, however, was found to be sensitive to scoring conditions (i.e., *top-ideas* and *time-on-task*). Reliability was lowest when only a single top-idea was considered, but it could be increased substantially by including some additional top-ideas (Benedek, Franz et al., 2012), and was highest for the average score which makes use of all ideas (Silvia et al. 2008). A straightforward explanation for this is that the aggregated evaluations of a larger number of ideas allows for a more reliable assessment, just as any test increases reliability by extending the number of relevant items. Also, considering more ideas could compensate for any discrepancies between the participants and the raters about what are considered to be the most creative ideas.

A higher time-on-task was found to increase reliability at least for scores using four or more top-ideas. This suggests that scoring a high number of ideas makes more sense when there is enough time for participants to generate large numbers of ideas. A task time of 2 or 3 minutes apparently already worked quite well for most scores; further increases in task time only added small increases in reliability.

We also examined correlations with other common indicators of creativity to estimate effects of task properties on the validity of the originality score. A priori, one could assume that the correlation pattern would generally match that of reliability as any lack of reliability necessarily impairs validity coefficients. Interestingly, this was not the case. While the reliability evidence of originality scores was highest for average scoring at 5 minutes time-on-task, the correlation with openness for this score was lowest. The highest validity coefficients were obtained for a task time of 2 minutes using a medium number of about 3 to 6 top-ideas. This raises the question why correlations did not increase with increasing number of top-ideas just as reliability did? It has to be remembered that people were instructed to generate as many unusual and creative ideas as possible. High creative people presumably were able to generate many unusual ideas, of which, however, only some are very creative and thus truly indicative of their potential for creative thought. Hence, when all ideas are considered, such as in the average scoring, the evaluations of more and less creative ideas become mixed up. This would result in a moderate total creativity score for a high creative person which could equally be attained by a less creative person who just generated a few moderately creative ideas. It hence can be concluded that subjective top-scoring may provide more valid scores than average scoring, even though the latter method may be somewhat more reliable in terms of internal consistency.

The question remains why the validity did not increase steadily with time-on-task like reliability did. This might be explained by the fact that originality generally increases over time (e.g., Beaty & Silvia, 2012; Mednick, 1962; Piers & Kirchner, 1971) but creative people overcome common ideas more quickly than less creative people (Benedek & Neubauer, under review). As a consequence, after a short time-on-task, creative people may already have come up with highly original ideas whereas less creative people have not. As the time-on-task proceeds less creative people eventually also come up with more creative

ideas, whereas high creative people can hardly further improve their performance to the same extent. Hence, the discernment between high and low creative people (i.e., validity) may be higher for shorter task times than for excessively long ones.

Originality showed significant correlations only with openness but not with self-reported ideational behavior. The absent significant correlations with ideational behavior suggest that the ideational behavior questionnaire is more indicative of ideational fluency (two sample items read “I come up with a lot of ideas or solutions to problems.” or ”I have always been an active thinker—I have lots of ideas.“; Runco et al., 2000).

Recommendations/Implications for Scoring of Divergent Thinking Tasks

Some straightforward recommendations concerning the adequate assessment of ideational fluency and originality can be derived from the results of this study. For ideational fluency, it appears to be quite simple to obtain a reliable and valid score. This can be achieved by using divergent thinking tasks with short task time of about two minutes. The originality score, however, appears to be more sensitive to task and scoring properties. First, originality scores were found to be more valid when using tasks with durations of about 2 to 3 minutes. This substantiates the common practice of using similar tasks durations. Using shorter or much longer tasks, however, might negatively affect the validity of scores. Second, the top-scoring method should consider a medium number of about three to six ideas. Using much fewer or much more ideas (e.g., Plucker et al., 2011) may result in less valid scores. Considering that using a higher number of top-ideas also implies that a higher total number of ideas has to be subjected to ratings, it could be a good compromise to use three top-ideas. For a time-on-task of 2 minutes participants generated on average ten ideas. Using only the three most creative ideas would help to reduce the rating effort by about 70% as compared to having to evaluate all ideas. Similar rates were reported by Silvia et al. (2008) for top-2 scoring. Moreover, more than 90% of participants generated three or more ideas within two minutes.

Some Limitations of This Study and Future Directions

Some limitations of this study need to be addressed. Time-on-task was varied as an experimental variable by analyzing the performance data available at different times within the task. This was done in order to estimate scores that could be obtained for tasks of different length. While this method is efficient, results obtained for e.g. a time-on-task of 2 minutes might not fully generalize to studies which explicitly use 2 minute tasks. Differences might for example relate to higher effects of fatigue, since performing six divergent thinking tasks with five minutes probably involves more cognitive effort than six tasks with only two minutes. Moreover, people might apply different idea generation strategies when they know that tasks are shorter. We tried, however, to minimize these effects by not telling participants about the exact task time and by not giving them any information about the remaining task time.

A similar argument applies to the experimental variation of the number of top-ideas. The post-hoc selection of a specific number of top-ideas may not fully generalize to the corresponding instruction to select a specific number of top-ideas. Some people who had

generated large amounts of ideas reported that they found it difficult to arrange them all properly for creativity. This issue might be less prominent for shorter tasks and when people are just asked to identify their three or five most creative ideas. Taken together, it could be assumed that shorter task durations and the selection of a low number of most creative ideas may cause lower fatigue and more accurate judgments, which might eventually have additional positive effects on the psychometric properties of the originality score. Further limitations include the sample size, and the specific tasks which were selected for this study. For example, for more complex divergent thinking tasks (e.g., Reiter-Palmon et al., 2009), which often show lower fluency, the most adequate number of top-ideas might differ. The present findings hence await replication with larger samples, using other divergent thinking tasks, and employing further criteria for examining the validity of scores.

There are also some additional methodological issues that could be addressed in future research. First of all, one might consider employing separate tasks for assessing fluency and originality. We derived both scores from the same tasks (e.g., Torrance, 2008) and instructed participants to generate as many unusual and creative ideas as possible. This could be considered as kind of a double task which permits participants to employ different strategies either focusing on fluency or creativity of ideas. Future work might therefore attempt to assess fluency and originality with separate tasks using specific task instructions to focus either only on fluency or only on originality of ideas. While this procedure may require a larger total number of tasks, it might help to further increase the validity of scores. Finally, it should be noted that using a specific number of top-ideas also implies the possibility that there are some participants who actually do not generate as much ideas within the given time. There are different ways to handle this. In this study, we then used all available ideas of the participant. Another possibility would be to assign missing ideas with the lowest possible creativity rating (i.e., a creativity rating of zero). This would implicitly penalize very low fluency. Some side analyses indicated that such originality scores can again be highly correlated with fluency, at least for high numbers of top-ideas. This scoring approach could, however, be useful in studies which decide not to use separate fluency scorings but still allow for a moderate influence of fluency on the originality score.

Conclusions

This study provides further evidence of the usefulness of the subjective top-scoring method for the assessment of ideational originality (cf., Silvia et al. 2008). Using subjective top-scoring ensures that ideational originality scores overcome the issues often associated with this score, such as a lack of discriminant validity with respect to fluency. Moreover, adequate scoring methods help to obtain a highly reliable and valid originality score. As an example, a top-3 originality score for 2 minutes time-on-task showed a higher correlation with openness than fluency did. Adequate scoring of ideational originality hence may provide researchers with a powerful indicator of creative potential, besides and beyond fluency.

Acknowledgments

This research was supported by a grant from the Austrian Science Fund (FWF): P23914.

Europe PMC Funders Author Manuscripts

References

Amabile T. Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology. 1982; 43:997–1013.

Beaty RE, Silvia PJ. Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. Psychology of Aesthetics, Creativity, and the Arts. 2012 Advance online publication. doi: 10.1037/a0029171.

Benedek M, Fink A, Neubauer A. Enhancement of ideational fluency by means of computer-based training. Creativity Research Journal. 2006; 18:317–328.

Benedek M, Franz F, Heene M, Neubauer AC. Differential effects of cognitive inhibition and intelligence on creativity. Personality and Individual Differences. 2012; 53:480–485. [PubMed: 22945970]

Benedek M, Könen T, Neubauer AC. Associative abilities underlying creativity. Psychology of Aesthetics, Creativity, and the Arts. 2012; 6:273–281.

Benedek M, Neubauer AC. Revisiting Mednick's model on creativity-related differences in associative hierarchies. Evidence for a common path to uncommon thought. Manuscript under review.

Borkenau, P.; Ostendorf, F. NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae [NEO-Five factor inventory after Costa and McCrae]. Hogrefe; Göttingen: 1993.

Clark PM, Mirels HL. Fluency as a pervasive element in the measurement of creativity. Journal of Educational Measurement. 1970; 7:83–86.

Fink A, Benedek M. EEG Alpha power and creative ideation. Neuroscience and Biobehavioral Reviews. (in press). Advance online publication. doi:10.1016/j.neubiorev.2012.12.002.

Gilhooly KJ, Fioratou E, Anthony SH, Wynn V. Divergent thinking: strategies and executive involvement in generating novel uses for familiar objects. British Journal of Psychology. 2007; 98:611–625. [PubMed: 17535464]

Guilford, JP. The nature of human intelligence. McGraw-Hill; New York: 1967.

Hocevar D. A comparison of statistical infrequency and subjective judgment as criteria in the measurement of originality. Journal of Personality Assessment. 1979a; 43:297–299. [PubMed: 16367009]

Hocevar D. Ideational fluency as a confounding factor in the measurement of originality. Journal of Educational Psychology. 1979b; 71:191–196.

Jauk, E.; Benedek, M.; Dunst, B.; Neubauer, AC. The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. Manuscript under review

Kaufman, JC.; Plucker, JA.; Baer, J. Essentials of Creativity Assessment. John Wiley & Sons; Hoboken, NJ: 2008.

Kuhn JT, Holling H. Measurement invariance of divergent thinking across gender, age, and school forms. European Journal of Psychological Assessment. 2009; 25:1–7.

Lau S, Cheung PC. Creativity assessment: Comparability of the electronic and paper-and-pencil versions of the Wallach–Kogan Creativity Tests. Thinking Skills and Creativity. 2010; 5:101–107.

Mednick SA. The associative basis of the creative process. Psychological Review. 1962; 69(3):220–232. [PubMed: 14472013]

Michael, WB.; Wright, CR. Psychometric issues in the assessment of creativity. In: Glover, JA.; Ronning, RR.; Reynolds, CR., editors. Handbook of Creativity. Plenum Press; New York: 1989. p. 33-75.

Mouchiroud C, Lubart T. Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. The Journal of Genetic Psychology. 2001; 162:382–401. [PubMed: 11831349]

Nusbaum EC, Silvia PJ. Are intelligence and creativity really so different? Fluid intelligence, executive processes, and strategy use in divergent thinking. Intelligence. 2011; 39:36–45.

Piers EV, Kirchner EP. Productivity and uniqueness in continued word association as a function of subject creativity and stimulus properties. Journal of Personality. 1971; 39(2):264–276. [PubMed: 5581827]

Plucker JA, Qian M, Wang S. Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. Journal of Creative Behavior. 2011; 45:1–22.

Reiter-Palmon R, Illies MY, Cross LK, Buboltz C, Nimps T. Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. Psychology of Aesthetics, Creativity, and the Arts. 2009; 3:73–80.

Runco MA. Maximal performance on divergent thinking tests by gifted, talented, and nongifted children. Psychology in the Schools. 1986; 23:308–315.

Runco MA. Commentary: Divergent thinking is not synonymous with creativity. Psychology of Aesthetics, Creativity, and the Arts. 2008; 2:93–96.

Runco MA, Acar S. Divergent thinking as an indicator of creative potential. Creativity Research Journal. 2012; 24:66–75.

Runco MA, Albert RS. The reliability and validity of ideational originality in the divergent thinking of academically gifted and nongifted children. Educational and Psychological Measurement. 1985; 45:483–501.

Runco MA, Mraz W. Scoring divergent thinking tests using total ideational output and a creativity index. Educational and Psychological Assessment. 1992; 52:213–221.

Runco MA, Okuda SM, Thurston BJ. The psychometric properties of four systems for scoring divergent thinking tests. Journal of Psychoeducational Assessment. 1987; 5:149–156.

Runco MA, Plucker JA, Lim W. Development and psychometric integrity of a measure of ideational behavior. Creativity Research Journal. 2000; 13:393–400.

Silvia PJ. Discernment and creativity: How well can people identify their most creative ideas? Psychology of Aesthetics, Creativity, and the Arts. 2008; 2:139–146.

Silvia PJ. Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. Thinking Skills and Creativity. 2011; 6:24–30.

Silvia PJ, Martin C, Nusbaum EC. A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. Thinking Skills and Creativity. 2009; 4:79–85.

Silvia PJ, Winterstein BP, Willse JT, Barona CM, Cram JT, Hess KI, et al. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts. 2008; 2:68–85.

Smith, SM.; Ward, TB.; Finke, RA. The creative cognition approach. MIT Press; Cambridge: 1995.

Torrance, EP. Torrance Tests of Creative Thinking: Norms-technical manual, verbal forms A and B. Scholastic Testing Service; Bensenville, IL: 1974.

Torrance, EP. Torrance Tests of Creative Thinking: Norms-technical manual, verbal forms A and B. Scholastic Testing Service; Bensenville, IL: 2008.

Velicer, WF.; Eaton, CA.; Fava, JL. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In: Goffin, RD.; Helmes, E., editors. Problems and solutions in human assessment. Kluwer; Boston: 2000. p. 41-71.

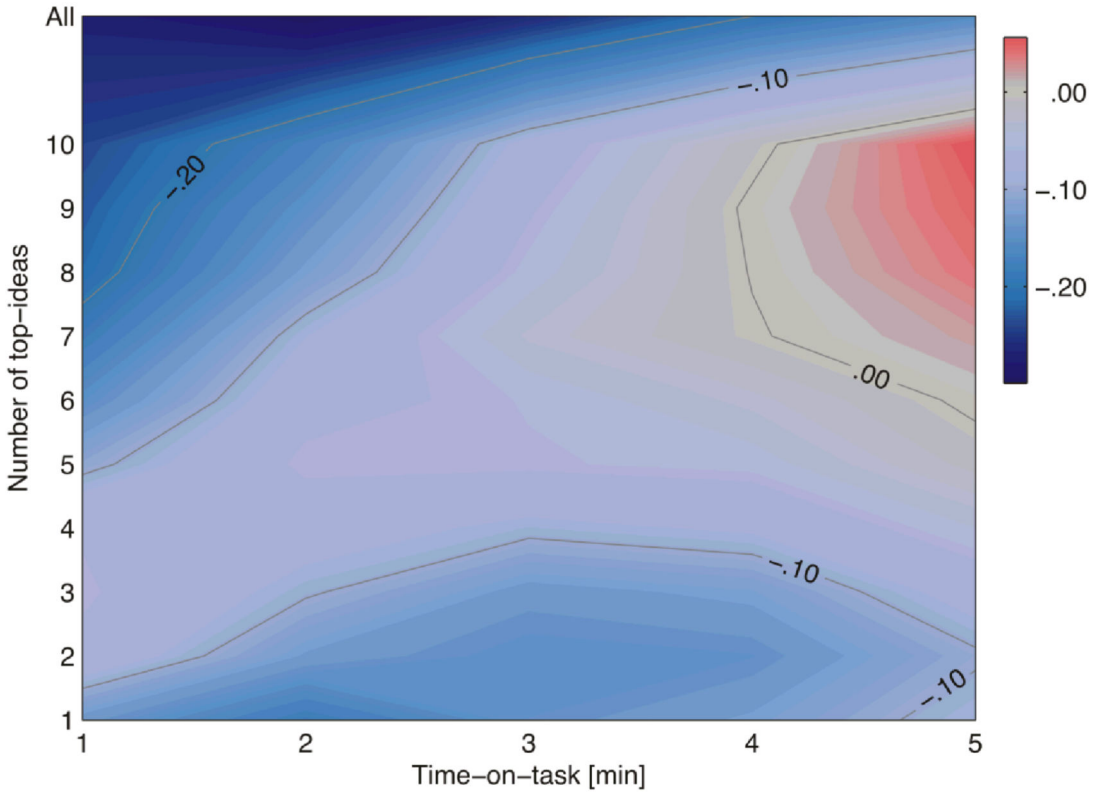

Figure 1.

Correlation of fluency and originality scores depending on the number of *top-ideas* and *time-on-task*. Correlation coefficients exceeding $r = .19$ are considered statistically significant given the sample size of $n = 105$.

Europe PMC Funders Author Manuscripts

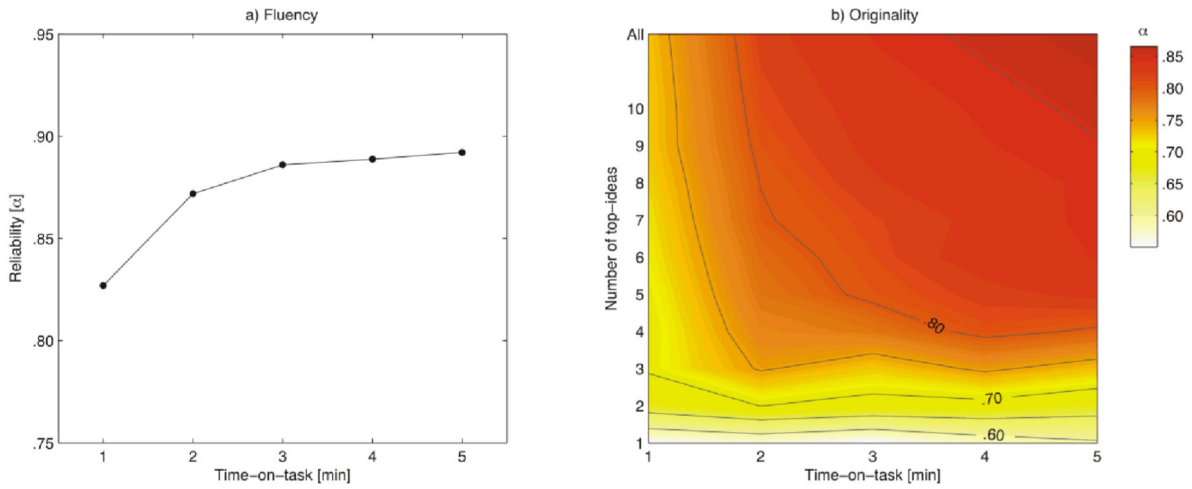

Figure 2.

Reliability (Cronbach's α) of a) the fluency score and b) the originality score depending on the number of *top-ideas* and *time-on-task.*

Europe PMC Funders Author Manuscripts

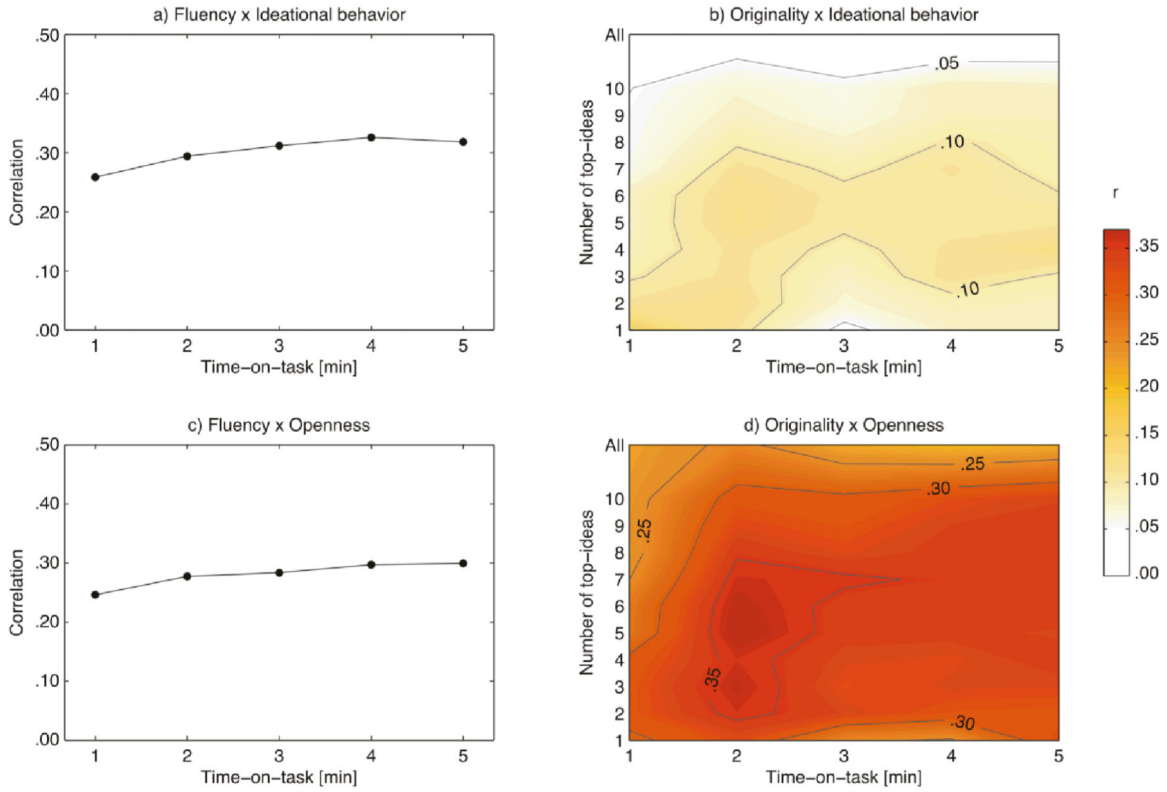

Figure 3.

Correlation of fluency (a, c) and originality (b, d) with self-reported ideational behavior and openness depending on the number of *top-ideas* and *time-on-task*. Correlation coefficients exceeding $r = .19$ are considered statistically significant given the sample size of $n = 105$.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Number of ideas generated after a time-on-task of 1 to 5 minutes in the alternate uses tasks and the instances tasks. The three values in each cell denote 10, 50, and 90 percentile values.

	1 min	2 min	3 min	4 min	5 min
Alternate uses	2.0/4.3/6.7	3.5/7.0/10.3	4.8/9.0/13.7	5.3/11.0/17.0	6.3/12.0/19.0
Instances	2.5/7.7/11.3	4.8/12.0/17.7	6.5/15.3/23.7	8.4/18.7/28.3	9.0/21.7/32.7