A New Perspective on the Multidimensionality of Divergent Thinking Tasks

Boris Forthmann[1]*, Paul-Christian Bürkner[1], Carsten Szardenings[2], Mathias Benedek[3] and Heinz Holling[1]

[1] Institut für Psychologie, University of Münster, Münster, Germany, [2] Fakultät Statistik, TU Dortmund University, Dortmund, Germany, [3] Institut für Psychologie, University of Graz, Graz, Austria

In the presented work, a shift of perspective with respect to the dimensionality of divergent thinking (DT) tasks is introduced moving from the question of multidimensionality across DT scores (i.e., fluency, flexibility, or originality) to the question of multidimensionality within one holistic score of DT performance (i.e., snapshot ratings of creative quality). We apply IRTree models to test whether unidimensionality assumptions hold in different task instructions for snapshot scoring of DT tests across Likert-scale points and varying levels of fluency. It was found that evidence for unidimensionality across scale points was stronger with be-creative instructions as compared to be-fluent instructions which suggests better psychometric quality of ratings when be-creative instructions are used. In addition, creative quality latent variables pertaining to low-fluency and high-fluency ideational pools shared around 50% of variance which suggests both strong overlap, and evidence for differentiation. The presented approach allows to further examine the psychometric quality of subjective ratings and to examine new questions with respect to within-item multidimensionality in DT.

Keywords: IRTrees, item-response theory, dimensionality, creativity, creative quality, fluency, divergent thinking

INTRODUCTION

Divergent thinking (DT) tasks are one of the most important proxies of creative thinking (Runco and Acar, 2012). For example, they are frequently used in research on the link of intelligence and creativity (e.g., Karwowski et al., 2016) and have been shown to predict creative achievement above intelligence (Kim, 2008). These tasks typically ask participants to come up with either many or creative ideas in order to solve a given problem. A classic research issue relating to such tasks is their underlying dimensionality; this has been discussed and researched since the 1950's. Most of the studies in this vein addressed the factorial validity of Guilford's classic four divergent-thinking abilities (i.e., fluency, flexibility, elaboration, and originality) in several common tests and batteries (e.g., Hargreaves and Bolton, 1972; Follman et al., 1973; Plass et al., 1974; Aliotti et al., 1975; Heausler and Thompson, 1988).

In the current work, however, we focus on the dimensionality of creative quality. In line with Forthmann et al. (2017a), we refer to creative quality scores as all scorings that have a clear definitional link to the concept of creativity. Prominently, originality scores fit this notion

of creative quality well, because originality is one of the defining characteristics of creativity (e.g., Runco and Jaeger, 2012). Thus, all commonly used indicators of originality, such as uncommonness, remoteness, and cleverness (see Wilson et al., 1953), are examples of creative quality scores (see Forthmann et al., 2017a). Moreover, this definition of creative quality allows including usefulness-based scores as it is the second defining characteristic of creativity (e.g., Runco and Jaeger, 2012). Beyond the two-componental standard definition of creativity, the definition of creative quality is open to other conceptualizations of creativity such as Simonton's three-componental definition (Simonton, 2012, 2018) or Kharkhurin's four-criterion definition (Kharkhurin, 2014), for example.

Various examples of quality scores can be found in the literature on DT. First, originality has been scored with respect to uncommonness (e.g., Runco and Charles, 1993; Mouchiroud and Lubart, 2001), remoteness (e.g., Christensen et al., 1957; Piers et al., 1960; Parnes, 1961), and cleverness (e.g., French et al., 1963; Mullins, 1963; Forthmann et al., 2017a). Second, usefulness has been scored independent of originality (e.g., Runco and Charles, 1993; Diedrich et al., 2015; Olteteanu and Falomir, 2016). Third, both components of the standard definition of creativity – originality and usefulness – were combined in an all-in-all creativity score (e.g., Runco and Mraz, 1992; Runco and Charles, 1993; Mouchiroud and Lubart, 2001; Storm and Patel, 2014; Diedrich et al., 2015). Clearly, there is no unanimous way to score DT responses for creative quality (see also Reiter-Palmon et al., 2019). In the current work, a creative quality score that is based on all classic indicators of originality will be used because this approach has been used more extensively over recent years (e.g., Silvia et al., 2008, 2009; Hass, 2015, 2017; Hofelich Mohr et al., 2016).

Subjective ratings with Likert-scales are often used to score creative quality (e.g., Silvia et al., 2008). In particular, providing subjective ratings for a person's ideational pool (all generated responses as a whole; see description below) will be in the focus of the current work (Runco and Mraz, 1992; Silvia et al., 2009). In fact, these ratings raise new questions with respect to dimensionality issues of DT tests. By a combination of ratings and a particular item-response modeling technique, so called IRTrees (De Boeck and Partchev, 2012), we show how to assess whether providing few or many ideas might induce multidimensionality in rated creative quality. Is there a relationship of quality in low-fluency sets and quality in high-fluency sets within the same person? Clearly, for the measurement of quality it is fundamental to know if the ability to be creative is being measured no matter how many ideas are generated. In fact, it is discussable if high-fluency and low-fluency ideational pools are unidimensional with respect to creative quality. Strategies that are efficient in terms of fluency (such as retrieval from memory) are most likely not conducive to idea quality which requires more demanding strategies (Gilhooly et al., 2007). Thus, differences in cognitive processes when people are fluent and when they are not may affect the dimensionality of creative quality scores.

In a similar vein, studies on the "be-creative" effect in DT are informative (Christensen et al., 1957; Harrington, 1975; Runco, 1986; Runco et al., 2005; Nusbaum et al., 2014; for

be-creative effects in other creative performance tasks see, for example, Chen et al., 2005; Niu and Liu, 2009; Rosen et al., 2017). In this line of research, standard instructions (be-fluent instructions) with a focus on quantity of responses (*think of as many ideas as possible*) are compared with explicit instructions to be creative (be-creative instructions; *think of ideas that are creative, original, uncommon, clever, and so forth*). Substantial variation in participants' strategies is expected, when the task goal remains opaque as in be-fluent instructions (Harrington, 1975; Nusbaum et al., 2014). On the contrary, "be-creative" instructions are assumed to homogenize participants' mindset toward the task and more demanding strategies can be expected to be used from the beginning (Nusbaum et al., 2014). Guilford (1968) argued also that explicit instructions to generate rather creative responses are likely to change the cognitive processes during idea generation. He expected a stronger involvement of evaluative processing with explicit instructions, which is in line with recent work by Nusbaum et al. (2014). As a consequence, the involvement of evaluative processing should be more homogeneous across participants when receiving explicit instructions to be creative as compared to be-fluent instructions. Thus, multidimensionality of creative quality of low-fluency and high-fluency ideational pools seem to be more likely under be-fluent instructions.

Moreover, the introduced modeling technique allows addressing another issue regarding the dimensionality of DT quality. This issue is related to the subjective scoring of DT quality. Subjective scorings of DT have gained prominence over the last years, but they were used also in the early years of DT research. One example of subjective scoring is the so-called snapshot scoring or scoring of ideational pools (Runco and Mraz, 1992; Silvia et al., 2009). Here raters see a full ideational pool of a participant and give a rating of the set's overall creative quality. These ratings are typically given on a 5-point Likert scale and here the question can be asked: Do we measure the same quality at the lower end and the upper end of the scale? Or to put it differently: Is one latent trait underlying the full rating scale? This question has been coined *ordinal hypothesis* by De Boeck and Partchev (2012).

An IRTree Modeling Approach for Divergent-Thinking Creative Quality Ratings

The model we use here is an IRT model for rating scales. It is inspired by the sequential model of Tutz (1990). In order to understand the model, it is helpful to use a linear response tree structure of the Likert scale (see **Figure 1**). A linear response tree is comprised of end nodes Y (i.e., the concrete response categories of the Likert scale) and internal nodes Y^* (De Boeck and Partchev, 2012). As shown in **Figure 1**, internal nodes can be understood as binary-coded sub-items. For example, the sub-item for the first internal node (Y_1^* in **Figure 1**) in the tree is coded zero when an ideational pool receives a rating of one, and the sub-item is coded one when the ideational pool received a rating larger than one. The sub-item for the second internal node (Y_2^* in **Figure 1**) in the tree is coded zero when an ideational pool

FIGURE 1 | Linear response tree for the creative quality ratings. The Y^*_r represent sub-items (De Boeck and Partchev, 2012) and the Y the response categories for the ratings.
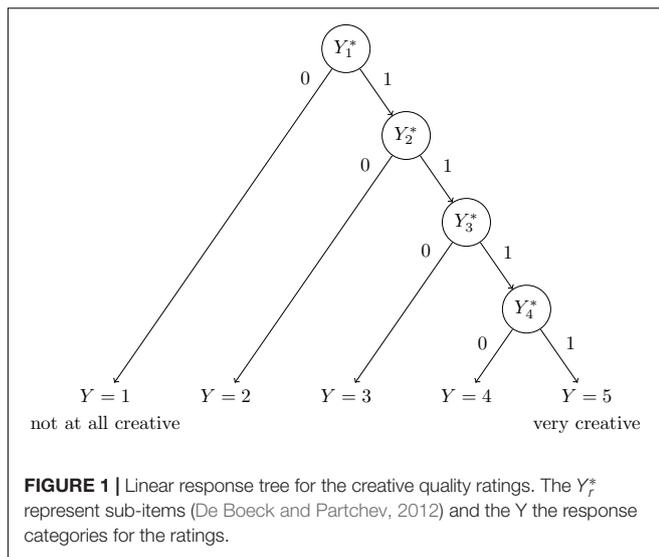
receives a rating of two, and the sub-item is coded one when the ideational pool received a rating larger than two, and so forth. Hence, the ratings on the Likert scale are divided into binary-coded sub-items Y^*_{pir} for each of the $r = 1, \ldots, R$ internal nodes. Please note that for simplicity we will use the term node from here onward to only refer to internal nodes of a response tree. Each node in the linear tree represents a branch with probability to remain at a given category $m = 1, \ldots, M - 1$ in contrast to all possible higher categories (De Boeck and Partchev, 2012). The model is based on the sub-items and can be formulated as:

$$P(Y^*_{pir} = y^*_{pir}|\theta_p) = \exp(\theta_{pr} + \beta_{ir})/[(1 + \exp(\theta_{pr} + \beta_{ir}))],$$

with $\theta_p = (\theta_{p1}, \ldots, \theta_{pr})$ being the node-specific latent variables (with person number $p = 1, \ldots, P$) and β_{ir} (with item number $i = 1, \ldots, I$) denoting a node-specific item threshold. For unidimensional models, only one latent ability is assumed to underlie the linear response tree. We will also use the term *node-unidimensionality* to refer to such unidimensional models. Moreover, as outlined in De Boeck and Partchev (2012), such a linear response tree can be combined with latent variable trees, which leads to other multidimensional modeling options. For example, a bi-factor structure can be proposed and tested as an underlying structure of a linear response tree.

Aim of the Current Study

The aim of the current study is to provide new insights into the dimensionality of DT by means of a proof of concept study using IRTree modeling. The IRTree modeling framework allows assessing dimensionality issues that are at the heart of frequently used scoring procedures in divergent-thinking assessment such as snapshot scoring of ideational pools (Runco and Mraz, 1992; Silvia et al., 2009). Related to this scoring procedure, we show how to test a unidimensional model underlying the rating scale against a multidimensional alternative that includes a different latent variable at each of the respective nodes. Unidimensionality of ratings in snapshot scoring with respect to the used scale is in

fact fundamental to the method, but it has not yet been assessed in the literature. It is important to know if the ability to score better as the lowest scale point is the same ability that allows a person to move from the second highest to the highest scale point of the response tree.

Moreover, by means of IRTree modeling, it is possible to take a different view on the relationship between fluency and creative quality ratings. We can assess if the latent dimensions underlying low-fluent and high-fluent creative quality are related or even tap to the same latent variable. This question was tested for standard and be-creative instructions. The goal of this research is to illustrate a methodological approach to new dimensionality issues of DT which moves from the question of multidimensionality across DT scores to the question of multidimensionality across the scale of DT scores and across ideational pools of varying size.

MATERIALS AND METHODS

Participants

The data set to illustrate the above ideas was taken from Forthmann et al. (2016). This dataset was gathered for a large project and used in previous publications on DT assessment issues (Forthmann et al., 2016, 2017a,b) and the relationship between multicultural experiences and DT performance (Forthmann et al., 2018). However, the analyses in this study are unique to this work and go beyond any of the issues tackled in the above-mentioned articles. In that study eight Alternate Uses Tasks (AUTs; Wallach and Kogan, 1965; Guilford, 1967) were administered online. Four of the tasks were administered with a be-fluent instruction and the other four with a be-creative instruction. The AUTs were scored for fluency and subjective ratings of creative quality of the full ideational pools were given by two raters. The raters were asked to give holistic ratings of creative quality reflecting the three classic originality indicators – uncommonness, remoteness, and cleverness – according to the rating scheme provided in Silvia et al. (2008): *Uncommonness: Any response that is given by a lot of people is common, by definition*; *remoteness: creative uses for an object are far from everyday uses and obvious responses*; and *cleverness: clever ideas strike people as insightful, ironic, humorous, fitting, or smart*. In addition, the coding scheme refers also to task appropriateness: *A random or inappropriate response would be uncommon but not creative*. For more details with respect to these scoring dimensions see Silvia et al. (2008; p. 85). The ratings were provided on a 5-point Likert-scale as depicted in **Figure 1**. Absolute agreement intra-class correlations for the average scores indicated good inter-rater reliability according to Cicchetti's criteria (Cicchetti, 2001), ICC(2,2) = 0.701, 95%-CI: [0.669, 0.743]. The final sample for data analysis was $N = 249$ (age: $M = 23.48$, $SD = 6.44$; gender: 79.12% female and 20.88% male). For more details please consult Forthmann et al. (2016).

This study was carried out in accordance with the recommendations for online studies provided by the ethics committee of the department of psychology of the University in Münster. These guidelines are in accordance with the

guidelines provided by the German foundation for online research (Deutsche Gesellschaft für Online Forschung). All subjects gave written informed consent prior to participation. An ethics approval was not required as per institutional and national guidelines.

Data Preparation and Statistical Analysis

The data were prepared with R functions provided by the IRTrees package (De Boeck and Partchev, 2012). In order to fit the proposed models, the function dendrify() was used to expand and reshape the data from a wide-format matrix into a long-format matrix of sub-item responses (see De Boeck and Partchev, 2012). The data frame that is built by this function includes an indicator variable for the nodes, so that unidimensionality vs. multidimensionality for the nodes can directly be tested.

Analogous to Partchev and De Boeck (2012) median splits were used in order to separate low-fluency ideational pools from high-fluency ideational pools. Moreover, a methodological sensitivity analysis was performed to assess the differences between median-splits (with respect to fluency) within items and within persons. For the within-item median split all medians for every AUT task were calculated separately for both instructions. An indicator variable for low-fluency vs. high-fluency ideational pools was then formed by a binary coding. The indicator was coded one if fluency for a given set was strictly below (instead of below or equals) the respective median since this led to a more balanced distribution of the indicator variable. The within-person split was analogous, but here the median values were calculated for each person and both instruction types separately. Based on these medians the same indicator variables were built. Moreover, possible differences between raters' central tendency of ratings were statistically controlled by including a corresponding fixed effect in all estimated models.

De Boeck and Partchev (2012) used maximum-likelihood methods to fit their models. For the present paper, however, we prefer a Bayesian approach for two reasons: First, random effects in the models refer to latent traits and latent trait correlations are of particular interest for our research question. Typical fitting procedures only return a point estimate for these correlations. Bayesian methods return the full posterior distribution of the correlations including standard errors and credible intervals (i.e., Bayesian confidence intervals). These provide a better understanding of the correlation structure and facilitate its interpretation. This is particularly relevant, if the posterior distribution of correlations is skewed, in which case the maximum-likelihood estimate is a poor measure of central tendency. Second, Bayesian methods have some advantages when it comes to model comparison. When using maximum-likelihood procedures, model comparisons are usually performed either by means of significance tests or by comparing information criteria. The former often lack statistical power when the number of observations is small, but for large data sets, even tiny and practically irrelevant differences in model fit will often get significant. These drawbacks are partially addressed by information criteria such as the AIC, but it remains unclear how much the criteria of two models should differ to indicate substantial differences in model fit. For recently developed

Bayesian specific information criteria, namely an approximation of the leave-one-out cross-validation (LOO), standard errors can be computed giving a much better sense of uncertainty in those criteria (Vehtari et al., 2017).

All statistical analyses were performed in R (R Core Team, 2013) using the brms package (Bürkner, 2016), which is based upon the probabilistic programming language Stan (Carpenter et al., 2017). It is further noteworthy, that refitting the models by means of the package lme4 (Bates et al., 2015) which uses restricted maximum likelihood estimation did not lead to relevant differences of the presented findings. The R code and data file are available in Open Science Framework[1].

RESULTS

Dimensionality Underlying the Nodes

A unidimensional linear-response tree model was compared with a model that assumed different latent variables for each of the corresponding nodes. This comparison was initially conducted for the full data set (thus, instruction as an influencing variable was ignored). This comparison was in favor of the multidimensional model as indicated by the difference in LOO criteria (see **Table 1**). When inspecting the correlation matrix of this multidimensional model it was apparent that latent variables referring to nodes of close proximity were more strongly correlated as compared to distant nodes. The three highest nodes correlated fairly strong with each other, whereas the lowest node correlated moderately ($r = 0.39$) with its next node and appeared to be unrelated to the other nodes. Thus, the ability to depart from the most common ideas (the ability to generate a response that receives a rating >1) appeared to be loosely related to the other abilities underlying the response tree. However, this pattern of results might be a reflection of different underlying strategies that either opt for quantity with the be-fluent instruction – participants are likely to accept low-quality ideas to increase fluency more readily here – or for quality with the be-creative instruction. Consequently, the dimensionality underlying the nodes was tested again for both instruction types separately.

For the instruction-specific tests of node-specific multidimensionality the item-parameters for the fourth node in the be-fluent instruction and the first node in the be-creative instruction were not identified due to lack of idea sets that received such ratings. According to the LOO information criterion, the multidimensional node model fit better in the be-fluent instruction, whereas for the be-creative instruction the unidimensional model was better (see **Table 1**). From the correlation estimates for latent variables of these models it is evident that support for unidimensionality of the nodes was stronger with a be-creative instruction as compared to the be-fluent instruction. The inter-correlations of latent variables underlying node 1, node 2, and node 3 ranged from $r = 0.37$ to $r = 0.76$ in the be-fluent instruction, whereas for the be-creative instruction the range was $r = 0.73$ to $r = 0.89$.

[1]https://osf.io/3rf24/

TABLE 1 | Comparison of unidimensional models vs. multidimensional models with respect to the nodes for the full model and for the be-creative instruction and be-fluent instruction.

	Node 1	Node 2	Node 3
Full model			
Node 2	0.38 [0.19, 0.57]	–	
Node 3	0.02 [−0.21, 0.26]	0.79 [0.65, 0.91]	–
Node 4	−0.13 [−0.50, −0.23]	0.56 [0.23, 0.81]	0.85 [0.59, 0.98]
Be-fluent			
Node 2	0.61 [0.37, 0.83]	–	
Node 3	0.35 [−0.11, 0.76]	0.77 [0.51, 0.96]	–
Node 4	NA	NA	NA
Be-creative			
Node 2	NA	–	
Node 3	NA	0.89 [0.74, 0.98]	–
Node 4	NA	0.73 [0.38, 0.96]	0.80 [0.52, 0.98]

	LOO (*SE*)		ΔLOO (*SE*)
Full model			
Unidimensional	8618.21 (113.09)		
Multidimensional	8397.68 (110.45)		220.52 (30.10)
Be-fluent			
Unidimensional	3742.83 (77.52)		
Multidimensional	3711.31 (75.90)		31.51 (13.23)
Be-creative			
Unidimensional	3882.10 (66.76)		
Multidimensional	3889.99 (67.33)		−7.89 (6.39)

n = 249. The unidimensional model implies the same latent variable for all nodes. LOO, information criterion approximating leave-one-out cross-validation; ΔLOO, LOO-difference between multidimensional and unidimensional models; NA, information in the data was insufficient for the highest node in the be-fluent instruction and the lowest node in the be-creative instruction.

Discriminating Creative Quality for Ideational Pools of High and Low Fluency

In order to account for insufficient information in the data at the first node in the be-creative instruction and the fourth node in the be-fluent instruction, those were again omitted. Then, separate analyses for both instructions were conducted (however, the interested reader will find results for the full dataset in the supplementary material or in the OSF see text footnote 1). Initially, the competing models were compared by means of the LOO criterion. With respect to the median splitting method, it was found that using within-person split indicators for high-fluency and low-fluency ideational pools performed better in terms of model fit as compared to the unidimensional model, and the within-item split indicators. Generally, modeling two latent variables for high vs. low-fluency sets increased model fit as compared to a unidimensional model (see **Tables 2**, **3**).

We then inspected the pattern of correlations and found estimates for within-item split of $r = 0.66$ (be-fluent; see **Table 2**) and $r = 0.73$ (be-creative; see **Table 3**) which implies that at least 44% of variation can be considered as common. Similarly, for the within-person split correlations of $r = 0.71$ (be-fluent; see **Table 2**) and $r = 0.70$ (be-creative;

TABLE 2 | Correlations of latent variables and model comparison results for a differentiation of low-fluency and high-fluency creative quality for be-fluent instructions only.

	High-fluency
Within-person split	
Low-fluency	0.69 (0.54, 0.83)
Within-item split	
Low-fluency	0.63 (0.33, 0.89)

	LOO (*SE*)
Model	
Unidimensional	3742.83 (77.52)
Within-person split	
Low-fluency vs. high-fluency	3693.82 (77.21)
ΔLOO (vs. unidimensional)	49.01 (15.50)
Within-item split	
Low-fluency vs. high-fluency	3725.43 (77.48)
ΔLOO (vs. unidimensional)	17.39 (10.32)

LOO, information criterion approximating leave-one-out cross-validation; ΔLOO, LOO-difference between multidimensional and unidimensional models.

TABLE 3 | Correlations of latent variables and model comparison results for a differentiation of low-fluency and high-fluency creative quality for be-creative instructions only.

	High-fluency
Within-person split	
Low-fluency	0.68 [0.53, 0.81]
Within-item split	
Low-fluency	0.72 [0.50, 0.90]

	LOO (*SE*)
Model	
Unidimensional	3882.10 (66.76)
Within-person split	
Low-fluency vs. high-fluency	3827.73 (67.30)
ΔLOO (vs. unidimensional)	54.37 (14.98)
Within-item split	
Low-fluency vs. high-fluency	3859.44 (66.98)
ΔLOO (vs. unidimensional)	22.66 (8.88)

LOO, information criterion approximating leave-one-out cross-validation; ΔLOO, LOO-difference between multidimensional and unidimensional models.

see **Table 3**) were found which implies an amount of shared variation of 49%. Thus, besides a strong construct overlap of creative quality for low-fluency and high-fluency ideational pools, there is also evidence for differentiation at the latent level. Moreover, the strong overlap of credible intervals for each of the correlations should be noted which indicates that the found slight differences between correlation estimates are negligible. Consequently, construct differentiation of low-fluency and high-fluency ideational pools with respect to creative quality latent variables can be considered to be similar for both instructions and both splitting methods (within-person or within-item).

DISCUSSION

The dimensionality of divergent-thinking indices such as originality and fluency has been an intriguing endeavor since the early years of creativity research. In the current study, we introduced IRTrees as an interesting model alternative to study the underlying dimensionality of DT. For example, we assessed the dimensionality of creative quality ratings over the rating scale in order to test the ordinal hypothesis of ratings (De Boeck and Partchev, 2012). In addition, within the IRTree approach, we shifted the focus from the relationship of quality and quantity and the corresponding debate on the lack of discriminant validity of fluency and originality toward a quality-quality relationship when creative quality is assessed for high-production and low-production ideational pools within two common important instructions. Altogether, this study represents a new perspective on the dimensionality of divergent-thinking scores away from classical factor analytic approaches.

Using subjective ratings is not free of criticism (see Runco, 2008) and testing the null hypothesis of node unidimensionality offers a way to empirically test how well they work and under which circumstances. For example, these findings highlight recent suggestions made by Reiter-Palmon et al. (2019) to carefully take instruction-scoring fit into account. Instruction-scoring fit refers to the congruence of instructions and scoring, for example, a be-creative instruction resonates well with a scoring of creative quality. While instruction-scoring fit is already preferable from a conceptual perspective, the findings of the current study highlight the importance of instruction-scoring fit because node-unidimensionality as a desirable psychometric property of subjective ratings was better supported when instructions and scoring were congruent. Thus, a Likert-type scale seems to work only in the intended way here if raters see responses of participants who were instructed to focus on quality rather than quantity. Most likely, this finding can be attributed to more consistent behavior as a consequence of a clear goal that is set by means of the instructions (e.g., Harrington, 1975; Nusbaum et al., 2014). Thus, this work presents a starting point to further investigate if unidimensionality underlying the rating scale can be maintained only for be-creative instructions, which should be tested in other samples (perhaps relying on a larger rater sample) in order to corroborate the observations here.

Tentatively, we conclude that subjective ratings of creative quality require instruction-scoring fit to be psychometrically sound. However, instruction-scoring fit should not be viewed as a dichotomy. For example, the original test material of the AUT from the Guilford group (Wilson et al., 1960) instructs participants to think of uses that are different from an object's most common use (which is also provided for each object during the test procedure). Thus, instruction-scoring fit for creative quality and the original Guilford instruction would be slightly higher as compared to the fully unrestricted be-fluent instruction used in the current study. Instruction-scoring fit could even be further enhanced when instructions to avoid the most common use (Wilson et al., 1960) are further supplemented by commonly generated example ideas which are also required to be avoided (Shin et al., 2018; George and Wiley, 2019). Hence, it is yet

unclear which degree of instruction-scoring fit (with explicit instructions to be creative representing maximum instruction-scoring fit when creative quality is of interest) is required for subjective ratings to be node-unidimensional.

In addition, with be-fluent instructions it is efficient to quickly retrieve as many ideas from memory as possible (similar to verbal fluency tasks; for a related discussion see Nusbaum et al., 2014). Idea sets resulting from such a strategy reflect production ability and are likely to receive rather low ratings on a 5-point scale. Otherwise some participants may still receive higher-ratings even when be-fluent instructions are given due to their need to be original, need for cognition, or other motivational state and in order to reach such higher ratings more demanding strategies are necessary. Consequently, such differences in underlying cognitive processes and strategies to generate idea sets of varying quality may have caused the observed lack of node-unidimensionality. In fact, from experimental psychology it is known that differences in cognitive processes may influence the dimensionality of latent variable models (Oberauer et al., 2005).

For measurement of creative quality, a unidimensional model that outperforms multidimensional alternatives is desirable. This seems to be particularly the case when participants receive a be-creative instruction. If participants look for imaginative ideas und do not focus on quantity, fluency can be conceived as a byproduct and differentiating for fluency should not affect unidimensionality of creative quality. However, it was illustrated that high-fluency and low-fluency ideational pools differ in terms of creative quality at the latent level (around 50% of the variance is shared). But how did this heterogeneity of latent traits emerge?

One potential explanation here is to assume that differential effects of practice, exhaustion, and current motivation affected the dimensionality here. First, it has been demonstrated that at late time points in DT different and more elaborate strategies are more likely to be used (Gilhooly et al., 2007; see also below). Thus, if eight tasks are administered such strategies can directly be used in later tasks if they are encountered/used while working on one of the previous tasks. Second, if participants get more and more exhausted over the course of testing, it is likely that cognitive resources are used in different ways. Similarly, Haager et al. (2016) showed that production tasks are perceived to be less and less interesting over the course of time, while at the same time performance drops. As a consequence, motivational states toward the end of a test session with multiple DT tasks are likely to be less beneficial for performance. Most likely, all of the above factors have affected dimensionality in the current sample. Furthermore, early positions of the objects during test administration in the current sample were more likely to result in fluency scores above the median and, consequently, indicators for high-fluency and low-fluency latent variables were unevenly coded across their position during administration.

Another explanation for the differentiation of creative quality between low-fluency and high-fluency ideational pools could be that raters judged the quality of pools with varying numbers of responses in a slightly different way. In fact, it has been demonstrated that the amount of information that needs to be judged has a detrimental effect on rater agreement when DT responses are rated for creative quality (Forthmann et al., 2017b).

Thus, it should not be overlooked that processing on the side of the raters might potentially influence the results here beyond the statistical control of rater severity effects as it was applied in the current study.

The differentiation between low-fluency and high-fluency ideational pools (reflecting differences in speed of the response generation) bears further interesting opportunities for the study of within-item multidimensionality. Ratings can be obtained for individual responses generated for the same item, and responses could also be grouped according to certain characteristics that can be assumed to influence within-item dimensionality. For example, a well-known phenomenon is the serial order effect (for example, Beaty and Silvia, 2012) in DT. It describes the tendency to generate ideas of better quality (for example, more original and remote ideas) toward the end of the allotted time on task. This effect points to differences in the underlying cognitive processes at the beginning and at the end of a test session. Consequently, it might be suggested that different abilities are involved over the time course and this hypothesis can be tested by the method outlined here.

Finally, it needs to be acknowledged that the current study is indeed limited to the creative quality scoring that was used, namely a holistic scoring of uncommonness, remoteness, and cleverness. Forthmann et al. (2017a) found strong correlations between such holistic ratings and ratings of the cleverness dimension only ($r = 0.82$ for latent variables), it is likely that subjective ratings for the cleverness dimension yield similar results as compared to the current work. It is, however, unclear if results generalize to other originality indicators, indicators of the usefulness component of the standard definition of creativity, or other indicators relating to components specific to alternative conceptions of creativity (e.g., Simonton, 2012, 2018; Kharkhurin, 2014). In relation to this, it should further be mentioned that a test of node-unidimensionality is a specific issue for subjective ratings relying on Likert-type scales, whereas the question of multidimensionality of creative quality as a function of the ideational pool size (low-fluency vs. high-fluency) might also be addressable for other ways to score creative quality (e.g., statistical frequency as an indicator of originality). In this regards, future studies are indeed needed to further expand our knowledge.

How to Detect and Deal With Technical Problems

In order to apply IRTrees in future studies, our work provides guidance on how to deal with possible technical problems such as not enough information at some of the nodes. Here it was demonstrated that the available information strongly depended on the instruction. That is, not enough information was available

for the fourth node with a be-fluent instruction and for the first node with a be-creative instruction. As a consequence, all available data for those nodes in instruction-specific models were excluded. In future attempts, the available information at all nodes should be initially checked. For example, it normally occurs that standard errors for fixed effects of non-informative nodes are unrealistically high. Furthermore, in Bayesian applications non-informative nodes can be detected by convergence problems and high correspondence of prior and posterior if proper priors are assigned to the respective parameters.

OVERALL CONCLUSION

In the current work, IRTrees were introduced as a method that allows a change of perspective on the multidimensionality of DT tasks. The methodology looks promising in order to further explore the psychometric quality of subjective ratings of DT ideational pools (which has been traditionally a hot topic in DT research). Moreover, new studies on within-item multidimensionality of DT and also multidimensionality due to practice, exhaustion, and motivation are promising future applications of the presented method.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations for online studies provided by the institutional ethics committee. These guidelines are in accordance with the guidelines provided by the German foundation for online research (Deutsche Gesellschaft für Online Forschung). All subjects gave informed consent prior to participation. An ethics approval was not required as per institutional and national guidelines.

AUTHOR CONTRIBUTIONS

BF designed the study, wrote the initial draft, analyzed the data, and revised the manuscript. P-CB analyzed the data and revised the manuscript. CS wrote the initial draft and revised the manuscript. MB revised the manuscript. HH designed the study and revised the manuscript.

FUNDING

REFERENCES

Aliotti, N. C., Britt, M. F., and Haskins, G. P. (1975). Relationships among creativity, intelligence, and achievement measures in upward bound students. *Psychol. Sch.* 12, 423–427. doi: 10.1002/1520-6807(197510)12:4<423::AID-PITS2310120409>3.0.CO;2-Y

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-8*. Available at: http://CRAN.R-project.org/package=lme4

Beaty, R. E., and Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychol. Aesthet. Creat. Arts* 6, 309–319. doi: 10.1037/a0029171

Bürkner, P. C. (2016). brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 20, 1–37. doi: 10.18637/jss.v076.i01

Chen, C., Kasof, J., Himsel, A., Dmitrieva, J., Dong, Q., and Xue, G. (2005). Effects of explicit instruction to "be creative" across domains and cultures. *J. Creat. Behav.* 39, 89–110. doi: 10.1002/j.2162-6057.2005.tb01252.x

Christensen, P. R., Guilford, J. P., and Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *J. Exp. Psychol.* 53, 82–88. doi: 10.1037/h0045461

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *J. Clin. Exp. Neuropsychol.* 23, 695–700. doi: 10.1076/jcen.23.5.695.1249

De Boeck, P., and Partchev, I. (2012). IRTrees: tree-based item response models of the GLMM family. *J. Stat. Softw.* 48, 1–28. doi: 10.18637/jss.v048.c01

Diedrich, J., Benedek, M., Jauk, E., and Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychol. Aesthet. Creat. Arts* 9, 35–40. doi: 10.1037/a0038688

Follman, J., Uprichard, E., Malone, W., and Coop, R. (1973). Multivariate analyses of the relationship between creativity and intelligence. *Paper presented at the annual meeting of the American Educational Research Association*, New Orleans.

Forthmann, B., Gerwig, A., Holling, H., C_elik, P., Storme, M., and Lubart, T. (2016). The be-creative effect in divergent thinking: the interplay of instruction and object frequency. *Intelligence* 57, 25–32. doi: 10.1016/j.intell.2016.03.005

Forthmann, B., Holling, H., Çelik, P., Storme, M., and Lubart, T. (2017a). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creat. Res. J.* 29, 257–269. doi: 10.1080/10400419.2017.1360059

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., et al. (2017b). Missing creativity: the effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Think. Skills Creat.* 23, 129–139. doi: 10.1016/j.tsc.2016.12.005

Forthmann, B., Regehr, S., Seidel, J., Holling, H., Çelik, P., Storme, M., et al. (2018). Revisiting the interactive effect of multicultural experiences and openness to experience on divergent thinking. *Int. J. Intercult. Relat.* 63, 135–143. doi: 10.1016/j.ijintrel.2017.10.002

French, J. W., Ekstrom, R. B., and Price, L. A. (1963). *Manual for Kit of Reference Tests for Cognitive Factors*. Princeton, NJ: Educational Testing Service.

George, T., and Wiley, J. (2019). Need something different? Here's what's been done: effects of examples and task instructions on creative originality. Manuscript submitted for publication.

Gilhooly, K. J., Fioratou, E., Anthony, S. H., and Wynn, V. (2007). Divergent thinking: strategies and executive involvement in generating novel uses for familiar objects. *Br. J. Psychol.* 98, 611–625. doi: 10.1111/j.2044-8295.2007.tb00467.x

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York, NY: McGraw-Hill.

Guilford, J. P. (1968). *Intelligence, creativity, and their educational implications*. San Diego, CA: Robert R. Knapp

Haager, J. S., Kuhbandner, C., and Pekrun, R. (2016). To be bored or not to be bored-how task-related boredom influences creative performance. *J. Creat. Behav.* 52, 297–304. doi: 10.1002/jocb.154

Hargreaves, D. J., and Bolton, N. (1972). Selecting creativity tests for use in research. *Br. J. Psychol.* 63, 451–462. doi: 10.1111/j.2044-8295.1972.tb01295.x

Harrington, D. M. (1975). Effects of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores. *J. Personal.* 43, 434–454. doi: 10.1111/j.1467-6494.1975.tb00715.x

Hass, R. W. (2015). Feasibility of online divergent thinking assessment. *Comput. Hum. Behav.* 46, 85–93. doi: 10.1016/j.chb.2014.12.056

Hass, R. W. (2017). Semantic search during divergent thinking. *Cognition* 166, 344–357. doi: 10.1016/j.cognition.2017.05.039

Heausler, N. L., and Thompson, B. (1988). Structure of the torrance tests of creative thinking. *Educ. Psychol. Measure.* 48, 463–468. doi: 10.1177/0013164488482021

Hofelich Mohr, A., Sell, A., and Lindsay, T. (2016). Thinking inside the box: visual design of the response box affects creative divergent thinking in an online survey. *Soc. Sci. Comput. Rev.* 34, 347–359. doi: 10.1177/0894439315588736

Karwowski, M., Dul, J., Gralewski, J., Jauk, E., Jankowska, D. M., Gajda, A., et al. (2016). Is creativity without intelligence possible? A necessary condition analysis. *Intelligence* 57, 105–117. doi: 10.1016/j.intell.2016.04.006

Kharkhurin, A. V. (2014). Creativity. 4in1: four-criterion construct of creativity. *Creat. Res. J.* 26, 338–352. doi: 10.1080/10400419.2014.929424

Kim, K. H. (2008). Meta-analyses of the relationship of creative achievement to both IQ and divergent thinking test scores. *J. Creat. Behav.* 42, 106–130. doi: 10.1002/j.2162-6057.2008.tb01290.x

Mouchiroud, C., and Lubart, T. (2001). Children's original thinking: an empirical examination of alternative measures derived from divergent thinking tasks. *J. Genet. Psychol.* 162, 382–401. doi: 10.1080/00221320109597491

Mullins, C. J. (1963). Prediction of creativity in a sample of research scientists. *IEEE Trans. Eng. Manage.* 10, 52–57. doi: 10.1109/TEM.1963.6446513

Niu, W., and Liu, D. (2009). Enhancing creativity: a comparison between effects of an indicative instruction "to be creative" and a more elaborate heuristic instruction on Chinese student creativity. *Psychol. Aesthet. Creat. Arts* 3, 93–98. doi: 10.1037/a0013660

Nusbaum, E. C., Silvia, P. J., and Beaty, R. E. (2014). Ready, set, create: what instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychol. Aesthet. Creat. Arts* 8, 423–432. doi: 10.1037/a0036549

Oberauer, K., Wilhelm, O., and Schmiedek, F. (2005). "Experimental strategies in multivariate research," in *Multivariate Research Strategies: Festschrift in Honor of Werner W. Wittmann*, eds A. Beauducel, B. Biehl, M. Bosnjak, W. Conrad, G. Schönberger, and D. Wagener (Aachen: Shaker Verlag), 119–149.

Olteteanu, A.-M., and Falomir, Z. (2016). Object replacement and object composition in a creative cognitive system. Towards a computational solver of the Alternative Uses Test. *Cognit. Syst. Res.* 39, 15–32. doi: 10.1016/j.cogsys.2015.12.011

Parnes, S. J. (1961). Effects of extended effort in creative problem solving. *J. Educ. Psychol.* 52, 117–122. doi: 10.1037/h0044650

Partchev, I., and De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence* 40, 23–32. doi: 10.1016/j.intell.2011.11.002

Piers, E. V., Daniels, J. M., and Quackenbush, J. F. (1960). The identification of creativity in adolescents. *J. Educ. Psychol.* 51, 346–351. doi: 10.1037/h0042669

Plass, H., Michael, J. J., and Michael, W. B. (1974). The factorial validity of the torrance tests of creative thinking for a sample of 111 sixth-grade children. *Educ. Psychol. Measure.* 34, 413–414. doi: 10.1177/001316447403400227

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reiter-Palmon, R., Forthmann, B., and Barbot, B. (2019). Scoring divergent thinking tests: a review and systematic framework. *Psychol. Aesthet. Creat. Arts.* 13, 144–152. doi: 10.1037/aca0000227

Rosen, D. S., Kim, Y. E., Mirman, D., and Kounios, J. (2017). All you need to do is ask? The exhortation to be creative improves creative performance more for nonexpert than expert jazz musicians. *Psychol. Aesthet. Creat. Arts* 11, 420–427. doi: 10.1037/aca0000087

Runco, M. A. (1986). Maximal performance on divergent thinking tests by gifted, talented, and nongifted children. *Psychol. Sch.* 23, 308–315. doi: 10.1002/1520-6807(198607)23:3<308::AID-PITS2310230313>3.0.CO;2-V

Runco, M. A. (2008). Commentary: divergent thinking is not synonymous with creativity. *Psychol. Aesthet. Creat. Arts* 2, 93–96. doi: 10.1037/1931-3896.2.2.93

Runco, M. A., and Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creat. Res. J.* 24, 66–75. doi: 10.1080/10400419.2012.652929

Runco, M. A., and Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personal. Ind. Diff.* 15, 537–546. doi: 10.1016/0191-8869(93)90337-3

Runco, M. A., Illies, J. J., and Reiter-Palmon, R. (2005). Explicit instructions to be creative and original: a comparison of strategies and criteria as targets with three types of divergent thinking tests. *Int. J. Creat. Prob. Solv.* 15, 5–15.

Runco, M. A., and Jaeger, G. J. (2012). The standard definition of creativity. *Creat. Res. J.* 24, 92–96. doi: 10.1080/10400419.2012.650092

Runco, M. A., and Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and creativity index. *Educ. Psychol. Measure.* 52, 213–221. doi: 10.1177/001316449205200126

Shin, H., Cotter, K. N., Christensen, A. P., and Silvia, P. J. (2018). Creative fixation is no laughing matter: the effects of funny and unfunny examples on humor production. *J. Creat. Behav.* doi: 10.1002/jocb.383

Silvia, P. J., Martin, C., and Nusbaum, E. C. (2009). A snapshot of creativity: evaluating a quick and simple method for assessing divergent thinking. *Think. Skills Creat.* 4, 79–85. doi: 10.1016/j.tsc.2009.06.005

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C., Cram, J., Hess, K. I., et al. (2008). Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychol. Aesthet. Creat. Arts* 2, 68–85. doi: 10.1037/1931-3896.2.2.68

Simonton, D. K. (2012). Taking the US patent office criteria seriously: a quantitative three-criterion creativity definition and its implications. *Creat. Res. J.* 24, 97–106. doi: 10.1080/10400419.2012.676974

Simonton, D. K. (2018). Defining creativity: don't we also need to define what is not creative? *J. Creat. Behav.* 52, 80–90. doi: 10.1002/jocb.137

Storm, B. C., and Patel, T. N. (2014). Forgetting as a consequence and enabler of creative thinking. *J. Exp. Psychol. Learn. Mem. Cognit.* 40, 1594–1609. doi: 10.1037/xlm0000006

Tutz, G. (1990). Sequential item response models with an ordered response. *Br. J. Mathemat. Stat. Psychol.* 43, 39–55. doi: 10.1111/j.2044-8317.1990.tb00925.x

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Wallach, M. A., and Kogan, N. (1965). *Modes of Thinking in Young Children: A Study of the Creativity-Intelligence Distinction.* New York, NY: Holt, Rinehart, & Winston.

Wilson, R. C., Christensen, P. R., Merrifield, P. R., and Guilford, J. P. (1960). *Alternate Uses–Form A: Manual of Administration, Scoring, and Interpretation,* 2nd Preliminary Edn. Beverly Hills, CA: Sheridan Supply Company.

Wilson, R. C., Guilford, J. P., and Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychol. Bull.* 50, 362–370. doi: 10.1037/h0060857