5-1-2019

Scoring divergent thinking tests: A review and systematic framework

Roni Reiter-Palmon

Boris Forthmann

Baptiste Barbot

Scoring Divergent Thinking Tests: A Review and Systematic Framework

Roni Reiter-Palmon University of Nebraska at Omaha

Boris Forthmann University of Münster

Baptiste Barbot Pace University and Yale University

Divergent thinking tests are often used in creativity research as measures of creative potential. However, measurement approaches across studies vary to a great extent. One facet of divergent thinking measurement that contributes strongly to differences across studies is the scoring of participants' responses. Most commonly, responses are scored for fluency, flexibility, and originality. However, even with respect to only one dimension (e.g., originality), scoring decisions vary extensively. In the current work, a systematic framework for practical scoring decisions was developed. Scoring dimensions, instructionscoring fit, adequacy of responses, objectivity (vs. subjectivity), level of scoring (response vs. ideational pool level), and the method of aggregation were identified as determining factors of divergent thinking test scoring. In addition, recommendations and guidelines for making these decisions and reporting the information in papers have been provided.

Keywords:

divergent thinking, assessment, fluency, originality, flexibility

Supplemental materials: http://dx.doi.org/10.1037/aca0000227.supp

While the field of creativity has managed to, for the most part, find agreement on the definition of creativity, researchers in this field are less clear on how creativity should be operationalized and measured. This issue has been at the forefront of creativity research from its early days, starting with Guilford's work on the structure of intellect (Guilford, 1967). The measurement of creativity is a nontrivial matter (Plucker & Renzulli, 1999), as reviews and previous research suggest that predictors or variables associated with creativity may be specific to the measure used (Hornberg & Reiter-Palmon, 2017; Plucker & Renzulli, 1999; Reiter-Palmon, Young Illies, Kobe Cross, Buboltz, & Nimps, 2009), questioning the external validity of creativity studies. The work of Guilford (1950) emphasized the study of creativity in everyday life and the importance of divergent thought for creative production. Since then, divergent thinking (DT) tasks have probably been the most often used measures in the field of creativity research (Plucker, Qian, & Wang, 2011; Reiter-Palmon & Tinio, 2018).

However, there are limited resources available to researchers interested in the various available scoring procedures of DT tests and lack of agreement on a uniform

way in which these scoring procedures should proceed (e.g., Plucker et al., 2011). In this article, the operationalization of and the various specific scores that can be derived from DT tests are discussed within a systematic framework to guide assessment choices. Finally, we suggest practical guidelines for DT test scoring whenever possible.

The Nature of DT and DT Tasks

Despite the frequent use of tests designed to measure DT in creativity research, they should not be conceived as measures of creativity per se (e.g., Guilford, 1966; Runco, 2008). DT is the ability to generate multiple solutions in response to a given stimulus or problem (Guilford, 1967), and therefore, DT tasks provide a measure of (mainly) capacity for idea generation. Idea generation is only one process of the many processes that make up the full creative process (e.g., Guilford, 1950; Reiter-Palmon, 2018). Thus, when DT is studied as one facet of the full creative process, we do so in isolation of other important facets such as problem finding, idea evaluation, and so forth. However, because idea generation is critical to creativity across domains, DT test scores can be understood as indicators of potential, and they have been found to be predictive of creative achievement (Guilford, 1966; Kim, 2008; Plucker, 1999; Runco, Millar, Acar, & Cramond, 2010). It is important to keep in mind, therefore, that DT is not a measure of the whole creativity phenomenon when planning a study or interpreting findings.

Further, DT tasks cannot (and should not) be used interchangeably as if they were a direct proxy of a general DT ability. Viewed from a purely psychometric standpoint, DT tasks suffer from a poor alternate-form reliability, estimated to be on the .30 –.40 range (e.g., Barbot, Besançon, & Lubart, 2016; see also Barbot, 2019). This reliability depends on a number of factors such as the domain of the task, the time given for resolution (greater individual differences appear with greater resolution time), prior experience with or salience of stimuli used (Forthmann, Gerwig, Holling, Çelik, Storme, & Lubart, 2016), or the nature of instructions (Harrington, 1975; Nusbaum, Silvia, & Beaty, 2014). This raises two important issues: the importance of relying on a range of DT tasks (rather than a single one) whenever possible and reporting accurately the nature of DT tasks involved. For example, Guilford's (1967) Structute of Intellect (SOI)-based taxonomy, which has received empirical support (Guilford, 1984), could be used to that end. Both of these issues must be addressed to avoid misinterpretation of findings or overgeneralization to "general DT" if not the whole creativity phenomenon.

Scoring of DT Tasks

Researchers have many choices when it comes to DT scoring, and scoring decisions have important bearing on the final scores they ultimately obtain and interpret (for an overview of the scoring process, see Figure 1). In particular, one must consider which DT response should enter the scoring process, which scoring dimension(s) should be used, whether the chosen dimension(s) should be scored subjectively by raters or by objective methods, and whether single responses should be scored and

then aggregated as opposed to scoring the ideational pool of each respondent. If using the former, which approach to aggregation should be then used? If using the latter, how should raters be instructed and what strategy should they use to mentally combine the impression across all responses? Together, important decisions on the scoring process should be made regarding six main issues: (1) dimension of DT to be scored, (2) instructions provided to the participants and their fit to the chosen scoring dimension, (3) adequacy of responses to be scored (or not), (4) method of scoring ("objective" vs. "subjective"), (5) unit of scoring (response level vs. ideational pool), and (6) how should scores be computed (e.g., method of aggregation of response-level scores).

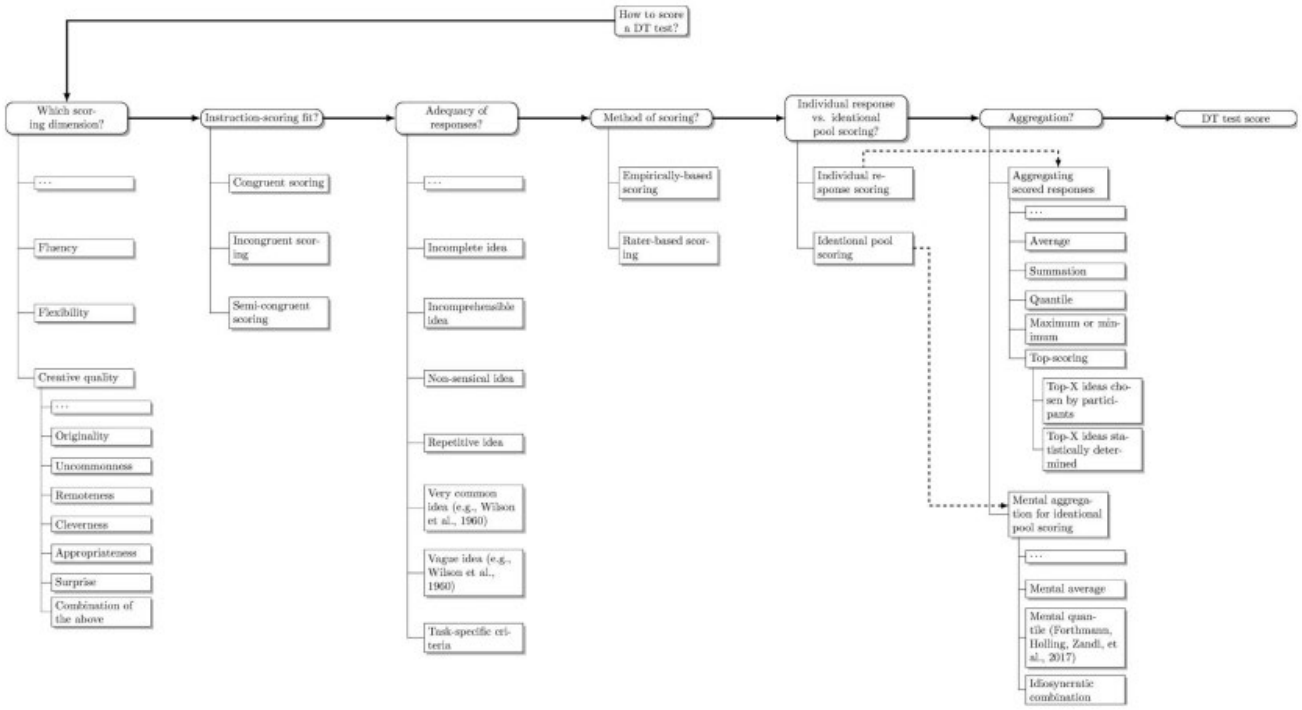

Figure 1. Systematic framework for the scoring of divergent thinking (DT) tests. Black arrows with a thick solid line characterize the required decision process to score DT tests. Solid lines lead to respective possible choices at each decision step. Black arrows with a dashed line characterize clear dependencies across decision steps. Centered dots imply options that are beyond the focus of the current article.

Choosing the DT Dimensions to Be Scored

The main dimensions of DT outlined by Guilford (1967) refer to both the quantity of ideas (fluency) and their quality, with respect to their diversity (flexibility), novelty (originality), and degree of elaboration. However, current research on DT tends to focus on fluency, followed by novelty and flexibility, with only a limited number of studies looking at elaboration (Hornberg & ReiterPalmon, 2017). For this reason, and for the sake of space saving, this article mainly focuses on fluency, flexibility, and originality.

Quantity of responses. A fluency score reflects the productivity aspect of DT. This dimension is the most often used in DT research and most consensually operationalized as a person's number of answers provided, in response to a given DT task. While some researches only count the number of adequate responses, some

include all responses, regardless of whether it actually addresses the prompt or task. In this respect, this dimension of DT is the least ambiguous to score. However, based on Guilford's original conceptualization of the construct of ideational fluency (*number of ideas per unit of time*, or "ideation rate") and recurrent concerns regarding response quality and fluency scores dependency, Barbot (2018) suggests that fluency may be best operationalized as the time required to generate a determined number of responses (e.g., how long does it take a test taker to generate 10 responses). This alternative approach allows one to standardize the number of DT outputs produced by each test taker, making comparisons of response quality more even. Regardless of its operationalization, fluency is sometimes used as the sole indicator of DT production. This approach has been criticized as the relationship between fluency and measures of creative achievement or performance is relatively low (Plucker et al., 2011), and fluency is not directly tied to the creativity definition of novelty and appropriateness. As a result, we recommend that an indicator of response quality always be used in conjunction with fluency.

Quality of responses. Originality is often reflected by the *uncommonness, remoteness*, and *cleverness* of ideas, whereas flexibility reflects the variety of responses (Guilford, 1967). Specifically, when evaluating responses for originality, the focus is on how unusual or uncommon the responses are, which can be determined using raterbased or frequency-based scoring methods. Both approaches along with methods of aggregation of response-level originality scores will yield their unique sets of challenges, such as fluency contamination effects (Plucker et al., 2011; see also below). To evaluate flexibility, the number of different categories of responses has typically been used. It has also been suggested that response quality can be assessed using a broader perspective aligned with the definition of creativity, which includes, for example, *originality, appropriateness,* or a combination of both (Forthmann, Holling, Çelik, Storme, & Lubart, 2017; Reiter-Palmon et al., 2009; Runco & Charles, 1993; Runco, Illies, & Reiter-Palmon, 2005). On this note, it is important to keep in mind that quality scorings may have a somewhat different meaning when solutions for more realistic problems are to be evaluated, indicating potential domain differences (Butler, Scherer, & Reiter-Palmon, 2003; Reiter-Palmon et al., 2009).

A recent alternative definition provided by Simonton (2018) includes a *surprise* component, which is tied to the prior knowledge of the person. Additionally, Simonton's definition provides a link to old/new scoring of responses (Gilhooly, Fioratou, Anthony, & Wynn, 2007). *Old/new* scoring instructs participants to classify their responses as "*old*" if they had the idea before working on the task (the idea is just retrieved; see also Runco & Acar, 2010) or "*new*" if they did not know the idea beforehand (the idea is generated rather than retrieved). The focus on this dimension is appropriate when there is an interest in memory processing underlying DT. This highlights the importance of theoretically grounded decisions when it comes to choosing a given dimension of DT to be scored, as opposed to reliance on mere conventions or ease of applicability.

Instruction-Scoring Fit

There is a large body of work on the effect of instructions on creative performance in DT measures (cf. Harrington, 1975; Runco et al., 2005), and sensitivity to slight variations of instructions has been critically appraised in studies on the response set bias of DT measures (Lissitz & Willhoft, 1985). The specific dimension(s) of DT of interest in a given study are optimally measured when instructions given are conceptually aligned. The rationale behind instruction-scoring fit is transparency of the task's goal. Only when the task goal is clear to participants can true individual differences in maximum performance with respect to the targeted dimension be observed (Harrington, 1975; Nusbaum et al., 2014). For example, participants can be assumed to propose responses that are creative, at least according to their own standards and understanding of the construct, only if they are instructed to be creative (Forthmann et al., 2016; Harrington, 1975; Nusbaum et al., 2014; Runco et al., 2005). Accordingly, when participants are instructed to generate many ideas, fluency increases and originality decreases. When participants are instructed to generate creative ideas, originality increases and fluency decreases (Nusbaum et al., 2014).

For many research questions, using instruction scoring congruence is most appropriate. Furthermore, when individual differences in fluency, flexibility, and originality are of primary interest, instruction-congruent measurement of all dimensions should be used. As a result, more tasks with corresponding congruent instructions are needed, which extends overall testing duration and demands on test takers.

However, there can be exceptions. In some cases, semicongruent or hybrid instructions (see Table 1 for an example) that emphasize multiple dimensions can be used and have been found to increase creative performance (Butler et al., 2003). Sternberg (2012) theorizes that creative individuals develop a need for uniqueness, which manifests as a habit to be original. This implies that their originality should show up in DT tasks even when they are not asked to do so. Thus, from this perspective, incongruent originality can be the focus of a study.

Adequacy of Responses

DT tests are often scored after manual exclusion of inadequate responses defined according to several criteria. For example, the test manual from Guilford's Alternate Uses Test (Wilson, Christensen, Merrifield, & Guilford, 1960) provides strict guidelines with respect to what qualifies as adequate response. Prescriptions preclude responses that apply to any object such as *selling, borrowing*, and the like. In addition, lists of common uses are provided to facilitate exclusion of such responses. However, nowadays, such ideas may remain in the response sample and be used in scoring (Abraham, 2016). Very often only incomplete, nonunderstandable, or nonsensical ideas and repetitions are excluded prior to scoring (see Forthmann et al., 2016). We suggest that criteria used to define response adequacy should be reported (even if responses were not evaluated for adequacy), and those criteria should be selected in line of the

purpose of the study. It is important to note that there is some overlap between the adequacy of the response and appropriateness (which is sometimes rated), such that those responses that are not adequate are also not appropriate. However, adequate responses can still have varying degrees of appropriateness.

Table 1
Illustration of Instruction-Scoring Fit and Congruent vs. Incongruent Scoring

Instruction type	Example instruction snippets[a]	Scoring			Example articles[b]
		Fluency	Flexibility	Creative quality/ originality	
Be-fluent	Give as many ideas as you can (Runco & Acar, 2010) Produce as many solutions . . . as you can think of (Harrington, 1975)	Congruent	Incongruent	Incongruent	Forthmann, Regehr, et al. (2018)
Be-flexible	Now we would like you to give as many different ideas as you can . . . be flexible . . . focus on variety (Runco & Okuda, 1991)	Incongruent	Congruent	Incongruent	Runco and Okuda (1991)
Be-creative/be-original	In this task it is important for you to be creative as possible (Forthmann et al., 2016) The goal is to come up with creative ideas (Nusbaum, Silvia, & Beaty, 2014)	Incongruent	Incongruent	Congruent	Forthmann, Regehr, et al. (2018)
Hybrid: Be-fluent AND be-creative	Focus on generating creative and unusual uses AND list as many other uses for the object cue as you can (Madore, Jing, & Schacter, 2016)	Partially congruent	Incongruent	Partially congruent	Madore et al. (2016)

[a] There is more information with respect to the instructions in the cited work along with the presented instruction snippets. [b] The suggested articles include the mentioned instruction in the respective column and application of all of the three scoring dimensions.

Method of Scoring

After deciding which dimension to score in light of the corresponding DT task instructions chosen, scoring may rely on either or both rater-based (i.e., often coined "subjective" methods) and empirically based methods (often coined "objective" methods). Obviously, objectivity is a desirable property of a measure (see Runco, 2008), but empirically based scorings of DT dimensions have their problems and, thus, rater-based scorings remain a useful option.

Empirically based scoring. Empirically based scoring of originality is usually based on the statistical frequencies of each response in the study sample, tapping into response *uncommonness* (Forthmann, Holling, Çelik, et al., 2017; Mouchiroud & Lubart, 2001). However, it is generally overlooked that the process of building the response occurrence table to calculate response frequencies often involves subjective decisions as to whether responses are actually the same as or different from each other (which is also a recurrent issue in flexibility scoring). Therefore, it is important to identify equivalent responses, and it has been acknowledged in the literature that this can be troublesome at times (Runco & Mraz, 1992). A best practice in this regard would be to base similarity evaluations on feature overlap (Maki, Krimsky, & Muñoz, 2006). Maki et al. (2006) demonstrated that feature overlap can be rated accurately, as suggested by early DT research using frequency tabulation of DT responses (Cropley, 1972). However, it is further important to take into account that some feature differences might

be irrelevant. For example, Wilson et al. (1960) outline that “saying that a milk carton can be used to ‘hold orange juice’ is not sufficiently different form ‘used to hold milk’” (p. 4). Thus, it is recommended to focus on feature differences resulting in functional differences for Alternative Uses Tasks (AUT) (or, depending on the task, on other aspects of similarity). Overall, it is surprising that the issue of how to judge response similarity in cross-tabulations for frequency-based originality scoring has rarely been described in detail and studied in the literature.

Another important consideration relates to what is considered an original response. Researchers have used continuous frequency-based scores (Forthmann, Holling, Çelik, et al., 2017; Mouchiroud & Lubart, 2001) or different thresholds to identify uncommon responses such as idiosyncratic responses, or those responses given by less than 1%, 5%, 10%, or even 20% of the study sample (Plucker, Qian, & Schmalensee, 2014). However, frequency estimates for scoring responses as “uncommon” or not are more accurate with larger sample sizes (see the online supplemental material). Confidence intervals around those frequency estimates are narrower (i.e., estimates more precise) for a larger sample than for a smaller sample, regardless of the threshold used.[1] Moreover, following the call for instructionscoring fit, it is possible that rather common responses are less frequently proposed than normal with explicit instructions to be creative and, paradoxically, may artificially appear to be more original (Forthmann, Holling, Çelik, et al., 2017). It is therefore recommended that empirically based scoring be used only when the sample size is sufficiently large (for more elaborate considerations with respect to this issue, see the online supplemental material).

Flexibility is commonly scored according to preexisting categories or a category system constructed ad hoc based on given data. Interscorer reliability in this context tends to be high, although objectivity is lower as compared to fluency scoring. One approach for scoring flexibility is counting the total number of different categories that have been utilized. A variant of this scoring is the number of category switches (Guilford, 1967; Nusbaum & Silvia, 2011). With its focus on category transitions instead of category assignment, this variant is a useful alternative, in particular for idea generation process studies.[2]

1 As such, using an increased threshold value to determine response uncommonness, as has been suggested, will not compensate for the imprecision related to sample size. For example, a unique response in a sample of 50 participants would yield a frequency estimate of 2% (one-sided Clopper-Pearson 95% CI [0.0, 9.1]). A response that occurs four times in a sample of 200 participants also yields a 2% frequency estimate (95% CI [0.0, 4.5]). While both scenarios result in a 2% estimate, corresponding CIs suggest that it can be used with confidence within the larger sample (the response likely not exceeds a 5% threshold for scoring originality; i.e., the CI does not cover the 5% threshold) and with reservation within the small sample (5% threshold is covered by the CI). See also online supplemental Figure S1.

2In this review, we are not referring to empirically based scoring methods relying on latent semantic analysis and related approaches. These are discussed in greater length in Acar and Runco (2019).

Rater-based scoring. Another approach to score for *uncommonness* involves human raters with sufficient experience with the response pool for the task at hand who judge on a Likert-type scale.[3] Gaining the required experience may involve self-guided familiarization with the generated responses or rater training prior to the final ratings, as well as familiarity with the creativity literature. The other two indicators of originality, *remoteness* and *cleverness*, can be scored in a similar fashion. When *remoteness* is assessed using rater-based scoring, only some facets of *remoteness* might be relevant for rater instructions (e.g., feature overlap vs. part-whole relations). For example, for the AUT, rating instructions might be mainly concerned with functional relations. Moreover, *cleverness* as an indicator of originality is solely scorable using rater-based methods emphasizing *imaginativeness, ingenuity, funniness*, and *cunning aptness* of responses (i.e., the defining terms of *cleverness* in creativity research). Examples of various rating scales used in scoring of DT tests are available in much of the research (Forthmann, Holling, Çelik, et al., 2017; Reiter-Palmon et al., 2009; Runco et al., 2005; Silvia et al., 2008).

In recent years, rater-based scorings have frequently used a combination of all three originality indicators: *uncommonness*, *remoteness*, and *cleverness* (Forthmann, Holling, Zandi, et al., 2017; Silvia et al., 2008). Indeed, Forthmann, Holling, Çelik, et al. (2017) found convergent validity evidence of several originality indicators, with only latent semantic analysis– based remoteness converging to a lesser extent with other scores. Furthermore, it should be noted that the combined rating scheme outlined in Silvia et al. (2008) may implicitly include the second component of creativity (i.e., appropriateness). That is, fuzzy, random, or vague ideas are set to receive low scores.

Relatedly, it is also possible to provide subjective ratings of *appropriateness* or *usefulness* for DT responses (e.g., Reiter-Palmon et al., 2009; Runco et al., 2005). Thus, with scorings for both facets of the standard definition of creativity, the conceptual connection of scoring and creativity would be highest. Further, combined ratings of both components have also been used for *overall creativity* ratings (Mouchiroud & Lubart, 2001). Finally, it should be noted that flexibility of ideational pools could rely on rater-based scoring using a Likert-type scale (ranging, e.g., from 1 *not flexible* at all to 5 *very flexible*), and the use of subjectively derived categories by the test takers themselves has been suggested (Snyder, Mitchell, Bossomaier, & Pallier, 2004).

Scoring of Individual Responses Versus Scoring of Ideational Pools

Scoring of the various dimensions of DT involves either considerations of individual responses or of the whole ideational pool. For instance, flexibility can only be scored in consideration of all responses provided by an individual, whereas originality of DT production can be scored at response level or at ideational pool level. Hence,

3 For discussion on specific practice and recommendations of procedure here, we refer the reader to the review on the consensual assessment technique by Cseh and Jeffries (2019).

decisions with respect to whether responses should be scored as a whole set or individually should be made, leading to follow-up decisions about how to aggregate response-level information.

Fluency scores are derived after scrutinizing each response's adequacy (Torrance, 1966; Wilson et al., 1960). In addition, flexibility or category switch scores require each response to be assigned into a conceptual category or to be marked as a conceptual switch or nonswitch as compared to the previously generated response (see Guilford, 1967). As suggested above, a subjective flexibility score could also be rated for full ideational pools. For originality, the question whether each response as opposed to the full ideational pool should be scored is naturally dependent on the choice between empirically based or rater-based scoring (see Figure 1).

With empirically based scorings, such as frequency-based originality, it is straightforward to score single responses prior to aggregation or statistical analyses, whereas for rater-based scorings, each ideational pool could also be scored as a whole ("snapshot scoring"; Forthmann, Holling, Zandi, et al., 2017; Plucker et al., 2014; Plucker et al., 2011; Runco & Mraz, 1992). While ideational pool scoring comes with the disadvantage that it prevents the response level to be the focus of analysis, it also reduces the overall burden on raters in terms of the number of ratings required.

However, while ideational pool scoring reduces the overall number of ratings, each single judgment is much more complex in terms of information processing. Forthmann, Holling, Zandi, et al. (2017) showed that this higher level of complexity is a likely source of rater disagreement and provide directions to reduce this complexity. Hass, Rivera, and Silvia (2018) found lower levels of reliability for ideational pool scoring compared to single idea ratings when laypersons were instructed to rate AUT and Consequences Task responses.

Finally, Silvia et al. (2008) introduced the notion of top-scoring. In this approach, participants select their top responses (most often two are chosen), those they believe are most creative, and only those are then rated. This, of course, reduces the amount of work associated with ratings. However, resulting scores reflect more complex thinking processes than just original idea generation, as it also supposes that test takers know to accurately select original responses (idea selection), which is not always the case (ReiterPalmon et al., 2009). In addition, researchers have also been concerned about the loss of information associated with the Top 2 scoring approach (e.g., Forthmann, Szardenings, & Holling, 2018). As only two ideas are scored for each individual, all the rest of the ideas are not scored or used.[4]

4 We thank an anonymous reviewer for raising that additional concern regarding the Top 2 scoring approach.

Aggregating Scored Responses

For fluency, aggregation is typically made by the summation of all adequate responses. For flexibility, the aggregated score results as the number of different assigned conceptual categories for a person's idea set. For category switching as a measure of flexibility, there are two scoring options. First, all switches are counted as one and then are summed (Guilford, 1967). The second approach counts switching only if the category has not been used before (Nusbaum & Silvia, 2011). However, any summation of weighted responses results in a confounding effect of flexibility scores by fluency (Forthmann, Szardenings, et al., 2018; Silvia et al., 2008). For correlational studies, these flexibility variants are best divided by fluency or, alternatively, residual scores can be used (Runco & Albert, 1985). However, for a study with a focus on mean comparison, flexibility scores based on summation can be meaningful (e.g., Forthmann, Regehr, et al., 2018).

The confounding effect of fluency in originality scores is often discussed (Forthmann, Szardenings, et al., 2018; Plucker et al., 2014; Plucker et al., 2011; Silvia et al., 2008). Forthmann, Szardenings, et al. (2018) provided a precise technical treatment of artifactual quality– quantity correlations, concluding that, for correlational studies, calculating ratio quality scores (whether originality, uncommonness, or another dimension) must be the default choice. Another reasonable alternative to ratio scores are residual scores (Runco & Albert, 1985), although this scoring potentially removes meaningful variation in originality beyond the artifactual overlap with fluency (Forthmann, Szardenings, et al., 2018). However, both average and residual scores potentially suffer from low reliability, which notably vary according to the correlation between fluency and the summative quality score (Arndt, Cohen, Alliger, Swayze, & Andreasen, 1991; Forthmann, Szardenings, et al., 2018). Overall, we have to admit that decisions here are heavily complex and affected by factors that cannot always be anticipated (e.g., the correlation between fluency and summative originality, which varies from study to study and heavily affects the reliability of ratios/average scores). For a more detailed treatment of issues around the effects of fluency contamination and potential solutions (e.g., ratio or residual scores), the reader should consult the above cited works. An illustration of possibilities to aggregate initially weighted ideas according to various scorings (beyond those mentioned above) can be found in online supplemental Table S1.

Practical Recommendations

Based on the discussion presented above, the following recommendations and guidelines are provided for decisions prior to data collection (design and methodology) and after data collection (scoring).

Prior to Data Collection

Prior to data collection, while designing the study, researchers should be careful to consider a number of issues. First is the type of DT task used. Depending on the purpose of the study, you may want to choose a specific domain and acknowledge that the task is not representative of all domains. However, if the purpose is to evaluate DT in general, then multiple domains and tasks must be included. Further, researchers must be careful here, as even within a domain, task types are not homogeneous (Reiter-Palmon et al., 2009). In addition, the dimensions on which DT scoring will be evaluated must also be considered at this point. This choice must be guided by the research questions posed and needs to occur early in the process, as it will influence the instructions given to participants. Specifically, depending on the purpose of the study and dimensions of interest, instruction-scoring congruence or hybrid instructions should be considered. As noted, in most cases, congruence would be a more appropriate choice. Importantly, if multiple dimensions of DT to be scored (e.g., fluency and originality) are to be used, and if congruent scoring is important, researchers may want to have separate tasks and instructions for fluency and separate ones for originality. This can quickly result in an increased number of tasks that the participant needs to engage in. Regardless of the final decision made by the researchers, DT studies should include information on the nature of the tasks (in that respect, we recommend using a SOI-based taxonomy) and the specific instructions that were used.

After Data Collection

Once the data have been collected, researchers must make decisions about scoring DT. The first issue is that of using response-level or response-set-level scoring, which will often be guided by prior decisions on dimensions of DT to be scored and other considerations such as sample size, availability of raters, and specific research questions. One important issue that emerges from using response-level scoring or scoring multiple dimensions of DT is the amount of work associated with such scoring. It is sometimes the reason that researchers choose to score fluency only, which is less ambiguous to score, and can be obtained quickly and easily. However, there are a number of ways that would allow researchers to obtain response-level quality scoring in a more efficient way (although this still takes time!).

By categorizing the responses first, researchers may be able to obtain response-level scoring more easily. As most research uses rater-derived categories, this is the process we will discuss. Two (or more) raters (blind to study conditions, if any) should review the full list of ideas, which would be presented in a random order. Each rater will then create a list of categories, with sample responses. At this point, one important consideration is the breadth of the categories. Broad categories will yield very few categories overall, whereas narrow categories will yield too many categories with very few responses— both of which will influence the flexibility score. The two raters should discuss and compare categories. Once each rater has created a list of categories, the raters will meet to compare and discuss categories to create a final list of categories. Once the list of categories is finalized, two raters will categorize the responses and

resolve any categorization differences. It is important that the category list and interrater agreement on categories (prior to discussion) be included when reporting results. Some practical issues to consider at this point: It is easiest to use an Excel spreadsheet for all of this work and note the category by each response. Second, a miscellaneous category should be created, and any response that cannot be categorized should be placed in this category. At that point, the responses in the miscellaneous category should be evaluated to determine if a new category has emerged.

Empirically based scoring includes scores on fluency, flexibility, and originality. Fluency is typically a direct count of ideas for each person. Here a decision needs to be made regarding whether ideas are redundant or not, as only nonredundant ideas are counted. More important, depending on the task and other scores, researchers may choose to count only relevant ideas, that is, those that fulfill the task requirements (as defined by the chosen response adequacy criteria, which must be precisely reported). However, if the dimensions include aspects of adequacy as part of the dimension definition (e.g., appropriateness), researchers may choose to still include these ideas but score them lower on that dimension. Flexibility scores can be obtained by summing the total number of different categories used or by evaluating switching as outlined above. The way in which a flexibility score was determined must be reported in the article. For empirically based originality scoring, the use of categories is particularly helpful. Similar responses will be placed in the same category, so it allows the researcher to view more easily how many of the same responses were provided in the data set and calculate a percent for each response.

Rater-based scoring for response-level data is also easier when responses are evaluated within a category. Raters can then provide the same rating more easily to the same response, as similar responses are grouped together. Although this is the dominant approach in rater-based scoring of originality, it is likely that it results in scores biased toward uncommonness. Indeed, raters may be implicitly influenced by the occurrence of identical responses in the pool of response they are judging. Therefore, an alternative procedure is to present raters with a “catalogue” of responses with all different responses in the sample presented only once, masking the occurrence of each response to the raters. The catalogue is then scored independently, and scores provided by each rater are matched back at the response level.

The final result here is a database that includes information at the response level for each participant. While fluency and flexibility yield one score per participant, the scoring for originality or other dimension, whether objective or subjective, results in multiple scores per participant and therefore should be aggregated, which was discussed in the previous section.

Discussion

The aim of the present review on DT scoring practices was to identify the major decisions related to scoring that researchers must engage in. We have identified

several facets of the scoring process, beginning with choice of scoring dimensions, scoring method (empirically based vs. rater based), level of scoring (idea vs. ideational pool level), and the method of aggregation, as critical decision points. We have further outlined several practical considerations with respect to determining adequacy and similarity of responses, which affects later steps of scoring procedures. In addition, we have provided practical recommendations whenever possible and identified gaps in the literature indicating the need for future research. These methodological differences may also account for some of the differences identified in the results of the many studies that use DT. However, our review indicates that different goals in research require different choices for scoring. Thus, it is expected that methodological setups in DT studies will to some degree remain different, but the outlined framework calls for a stronger theoretical justification with respect to measurement choices, which will strengthen future research.

As a final note, the outlined systematic framework will also facilitate studies on the robustness of findings in relation to DT. For example, the correlation between openness to experience with DT can be considered a benchmark finding because it was reported for several of the different choices outlined in this work. We argue that variety in DT assessment approaches should be used to reveal further benchmark findings, which are very useful when it comes to comparisons of competitive theories or the examination of new research topics related to DT.

References

Abraham, A. (2016). Commentary: Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. Frontiers in Psychology, 7, 824. http://dx.doi.org/10.3389/fpsyg.2016.00824

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. Psychology of Aesthetics, Creativity, and the Arts, 13, 153–158. http://dx.doi.org/10.1037/aca0000231

Arndt, S., Cohen, G., Alliger, R. J., Swayze, V. W., II, & Andreasen, N. C. (1991). Problems with ratio and proportion measures of imaged cerebral structures. Psychiatry Research, 40, 79 – 89. http://dx.doi.org/10.1016/0925-4927(91)90031-K

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. Frontiers in Psychology, 9, 2529. http://dx.doi.org/10.3389/fpsyg.2018.02529

Barbot, B. (2019). Measuring creativity change and development. Psychology of Aesthetics, Creativity, and the Arts, 13, 203–210. http://dx.doi.org/10.1037/aca0000232

Barbot, B., Besançon, M., & Lubart, T. (2016). The generality-specificity of creativity: Exploring the structure of creative potential with EPoC. Learning and Individual Differences, 52, 178 –187. http://dx.doi.org/10.1016/j.lindif.2016.06.005

Butler, A. B., Scherer, L. L., & Reiter-Palmon, R. (2003). Effects of solution elicitation aids and need for cognition on the generation of solutions to ill-structured problems. Creativity Research Journal, 15, 235–244. http://dx.doi.org/10.1207/S15326934CRJ152&3_13

Cropley, A. J. (1972). Originality scores under timed and untimed conditions. Australian Journal of Psychology, 24, 31–36. http://dx.doi.org/10.1080/00049537208255782

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. Psychology of Aesthetics, Creativity, and the Arts, 13, 159 –166. http://dx.doi.org/10.1037/aca0000220

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

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. Creativity Research Journal, 29, 257–269. http://dx.doi.org/10.1080/10400419.2017.1360059

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. Thinking Skills and Creativity, 23, 129 –139. http://dx.doi.org/10.1016/j.tsc.2016.12.005

Forthmann, B., Regehr, S., Seidel, J., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2018). Revisiting the interactive effect of multicultural experience and openness to experience on divergent thinking. International Journal of Intercultural Relations, 63, 135–143. http://dx.doi.org/10.1016/j.ijintrel.2017.10.002

Forthmann, B., Szardenings, C., & Holling, H. (2018). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. Psychology of Aesthetics, Creativity, and the Arts. Advance online publication. http://dx.doi.org/10.1037/aca0000196

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. British Journal of Psychology, 98, 611– 625. http://dx.doi.org/10.1111/j.2044-8295.2007.tb00467.x

Guilford, J. P. (1950). Creativity. American Psychologist, 5, 444 – 454. http://dx.doi.org/10.1037/h0063487 Guilford, J. P. (1966). Measurement and

creativity. Theory Into Practice, 5, 185–189. http://dx.doi.org/10.1080/00405846609542023

Guilford, J. P. (1967). The nature of human intelligence. New York, NY: McGraw-Hill.

Guilford, J. P. (1984). Varieties of divergent production. Journal of Creative Behavior, 18, 1–10. http://dx.doi.org/10.1002/j.2162-6057.1984.tb00984.x

Harrington, D. M. (1975). Effects of explicit instructions to “be creative” on the psychological meaning of divergent thinking test scores. Journal of Personality, 43, 434 – 454. http://dx.doi.org/10.1111/j.1467-6494.1975.tb00715.x

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. Frontiers in Psychology, 9, 1343. http://dx.doi.org/10.3389/fpsyg.2018.01343

Hornberg, J., & Reiter-Palmon, R. (2017). Creativity and the big five personality traits: Is the relationship dependent on the creativity measure? In G. Feist, R. Reiter-Palmon, & J. Kaufman (Eds.), The Cambridge handbook of creativity and personality research (pp. 275–293). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/9781316228036.015

Kim, K. H. (2008). Meta-analyses of the relationship of creative achievement to both IQ and divergent thinking test scores. Journal of Creative Behavior, 42, 106 –130. http://dx.doi.org/10.1002/j.2162-6057.2008.tb01290.x

Lissitz, R. W., & Willhoft, J. L. (1985). A methodological study of the Torrance Tests of Creativity. Journal of Educational Measurement, 22, 1–11. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01044.x

Madore, K. P., Jing, H. G., & Schacter, D. L. (2016). Divergent creative thinking in young and older adults: Extending the effects of an episodic specificity induction. Memory & Cognition, 44, 974 –988. http://dx.doi.org/10.3758/s13421-016-0605-z

Maki, W. S., Krimsky, M., & Muñoz, S. (2006). An efficient method for estimating semantic similarity based on feature overlap: Reliability and validity of semantic feature ratings. Behavior Research Methods, 38, 153–157. http://dx.doi.org/10.3758/BF03192761

Mouchiroud, C., & Lubart, T. (2001). Children’s original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. Journal of Genetic Psychology: Research and Theory on Human Development, 162, 382– 401. http://dx.doi.org/10.1080/002 21320109597491

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different? Fluid intelligence, executive processes, and strategy use in divergent thinking. Intelligence, 39, 36 – 45. http://dx.doi.org/10.1016/j.intell.2010.11.002

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to “be creative” reveals about the meaning and mechanisms of divergent thinking. Psychology of Aesthetics, Creativity, and the Arts, 8, 423– 432. http://dx.doi.org/10.1037/a0036549

Plucker, J. A. (1999). Is the proof in the pudding? Reanalyses of Torrance’s (1958 to present) longitudinal data. Creativity Research Journal, 12, 103–114. http://dx.doi.org/10.1207/s15326934crj1202_3

Plucker, J. A., Qian, M., & Schmalensee, S. L. (2014). Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking. Creativity Research Journal, 26, 135– 143. http://dx.doi.org/10.1080/10400419.2014.901023

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. Journal of Creative Behavior, 45, 1–22. http://dx.doi.org/10.1002/j.2162-6057.2011.tb01081.x

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

Reiter-Palmon, R. (2018). Creative cognition at the individual and team level: What happens before and after idea generation. In R. Sternberg & J. Kaufman (Eds.), The nature of human creativity (pp. 184 –208). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/9781108185936.015

Reiter-Palmon, R., & Tinio, P. P. (2018). 12 Years of PACA: A review of trends in PACA publications. Psychology of Aesthetics, Creativity, and the Arts, 12, 123–124. http://dx.doi.org/10.1037/aca0000185

Reiter-Palmon, R., Young Illies, M., Kobe Cross, L., Buboltz, C., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. Psychology of Aesthetics, Creativity, and the Arts, 3, 73– 80. http://dx.doi.org/10.1037/a0013410

Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. Psychology of Aesthetics, Creativity, and the Arts, 2, 93–96. http://dx.doi.org/10.1037/1931-3896.2.2.93

Runco, M. A., & Acar, S. (2010). Do tests of divergent thinking have an experiential bias? Psychology of Aesthetics, Creativity, and the Arts, 4, 144 –148. http://dx.doi.org/10.1037/a0018969

Runco, M. A., & Albert, R. S. (1985). The reliability and validity of ideational originality in the divergent thinking of academically gifted and nongifted children. Educational

and Psychological Measurement, 45, 483–501. http://dx.doi.org/10.1177/0013164485045003006

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. Personality and Individual Differences, 15, 537–546. http://dx.doi.org/10.1016/0191-8869(93)90337-3

Runco, M. A., Illies, J. J., & 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. Korean Journal of Thinking & Problem Solving, 15, 5–15.

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance tests of creative thinking as predictors of personal and public achievement: A fifty-year follow-up. Creativity Research Journal, 22, 361–368. http://dx.doi.org/10.1080/10400419.2010.523393

Runco, M. A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. Educational and Psychological Measurement, 52, 213–221. http://dx.doi.org/10.1177/0013164492052001026

Runco, M. A., & Okuda, S. M. (1991). The instructional enhancement of the flexibility and originality scores of divergent thinking tests. Applied Cognitive Psychology, 5, 435– 441. http://dx.doi.org/10.1002/acp.2350050505

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2, 68 – 85. http://dx.doi.org/10.1037/1931-3896.2.2.68

Simonton, D. K. (2018). Defining creativity: Don't we also need to define what is not creative? Journal of Creative Behavior, 52, 80 –90. http://dx.doi.org/10.1002/jocb.137

Snyder, A., Mitchell, J., Bossomaier, T., & Pallier, G. (2004). The creativity quotient: An objective scoring of ideational fluency. Creativity Research Journal, 16, 415–419. http://dx.doi.org/10.1080/1040041040 9534552

Sternberg, R. J. (2012). The assessment of creativity: An investment-based approach. Creativity Research Journal, 24, 3–12. http://dx.doi.org/10.1080/10400419.2012.652925

Torrance, E. P. (1966). Torrance tests of creative thinking. Normstechnical manual (Research ed.). Princeton, NJ: Personnel Press, Inc.

Wilson, R. C., Christensen, P. R., Merrifield, P. R., & Guilford, J. P. (1960). Alternate Uses–Form A: Manual of administration, scoring, and interpretation (2nd preliminary ed.). Beverly Hills, CA: Sheridan Supply Company.