5-1-2019

Creativity assessment in psychological research: (Re)setting the standards

Baptiste Barbot

Richard W. Hass

Roni Reiter-Palmon

Creativity Assessment in Psychological Research: (Re)Setting the Standards

Baptiste Barbot Pace University and Yale University

Richard W. Hass Thomas Jefferson University

Roni Reiter-Palmon University of Nebraska at Omaha

This commentary discusses common relevant themes that have been highlighted across contributions in this special issue on “Creativity Assessment: Pitfalls, Solutions, and Standards.” We first highlight the challenges of operationalizing creativity through the use of a range of measurement approaches that are simply not tapping into the same aspect of creativity. We then discuss pitfalls and challenges of the three most popular measurement methods employed in the field, namely divergent thinking tasks, productbased assessment using the consensual assessment techniques, and self-report methodology. Finally, we point to two imperative standards that emerged across contributions in this collection of articles, namely transparency (need to accurately define, operationalize, and report on the specific aspect[s] of creativity studied) and homogenization of creativity assessment (identification and consistent use of an optimal “standard” measure for each major aspect of creativity). We conclude by providing directions on how the creativity research community and the field can meet these standards.

Keywords:

creativity measurement, creativity assessment, validity, reliability, standards of assessment

The question of creativity measurement has often been at the forefront of the creativity research agenda (Plucker & Runco, 1998) since the beginning of the systematic study of individual differences in creativity, tracing back to over a century ago (e.g., Binet, 1900; Galton, 1869). Although some measurement methods have long been described as “gold standards” for creativity assessment— such as the Torrance Tests of Creative Thinking (TTCT; Torrance, 1998) or the Consensual Assessment Technique (CAT; Amabile, 1982)—persistent criticisms (e.g., Baer, 2011), evolving research needs (e.g., neuroscience paradigms; Benedek, Christensen, Fink, & Beaty, 2019), and efforts to address measurement challenges (e.g., Reiter-Palmon, Forthmann, & Barbot, 2019) have led to a plethora of alternative approaches to classic measurement methods and, in turn, a lack of set standards in creativity assessment. This lack of standards, together with enduring conceptual and methodological issues, greatly challenges the external validity of creativity studies, limits meta-analytical work, and makes the concept of creativity elusive to the novice eye.

This special issue provides a much-needed critical review of current practice in creativity assessment and existing measures, outlining common pitfalls while suggesting important guidelines and standards for best practice in creativity research and directions for the field. Such an endeavor is timely because we seem to be at an important pivotal point in the history of both creativity research and assessment. On the one hand, the increasing awareness of creativity as a relevant concept that can be measured and nurtured is evident from large-scale initiatives putting more than ever creativity "on the map." These initiatives include the Partnership for 21st Century Learning (P21; Partnership for 21st Century Skills, 2008) and the inclusion of *creative thinking* as part of the Organisation for Economic Co-operation and Development's Program for International Student Assessment (PISA)[1] 2021 program that, if moving forward, will assess half a million 15-year-old students worldwide. On the other hand, there is generally a continued interest in assessment issues (Plucker & Makel, 2010; Vartanian et al., 2019), which transpires from the increasing number of publications in (and submissions to) creativity research outlets. Indeed, although methods have not really changed for decades, as illustrated by Snyder, Hammond, Grohman, & KatzBuonincontro's (2019) review and Glaveanu's (2019) analysis, the field is now evolving rapidly. These changes are triggered notably through inputs from the neuroscience of creativity (Benedek, Christensen, Fink, & Beaty, 2019), the emergence of digital assessments and new assessment paradigms (Barbot, 2018b; Hass, 2017; Loesche, Goslin, & Bugmann, 2018), and new statistical development and approaches (e.g., Fürst, 2018; Myszkowski & Storme, 2019; Primi, Silvia, Jauk, & Benedek, 2019). These developments have provided new solutions and interpretations to enduring measurement problems (e.g., Forthmann, Szardenings, & Holling, 2018; Hornberg & Reiter-Palmon, 2017) and spurred renewed interest in the dynamization of creativity assessments (e.g., Jankowska, Czerwonka, Lebuda, & Karwowski, 2018). This is, in sum, an exciting time for creativity assessment, and this collection of articles provides a good snapshot of why it is so.

Which Creativity Construct Are We Measuring?

A Taxonomy of Creativity Assessment

In a recent commentary, Rossiter (2018) posited that "psychology— just as is standard practice in the physical sciences—needs to agree on an *optimal standard measure* of each major construct" (p. 931). Although creativity certainly qualifies as such a major construct, this noble objective might be unrealistic in our field. It is not that creativity researchers cannot agree in general. They have overall agreed on a standard, or general definition, of *creativity*. However, despite this consensus, the way creativity is operationalized from study to study greatly varies. The problem lies mainly in the broadness of the construct and the range of behaviors that are referred to by the

1 http://www.oecd.org/pisa/.

generic term *creativity*. Reaching consensus on an optimal standard measure of any psychological construct may work for relatively unified (i.e., unidimensional, and domaingeneral) and/or narrow (e.g., life-satisfaction) constructs but not for broad, multidimensional, and partly domain-specific phenomena such as creativity. In practice, there are many ways to operationalize creativity, but different creativity measures cannot be used interchangeably (e.g., Glaveanu, 2019; Hornberg & ReiterPalmon, 2017; Sternberg, 2018), because they all tap into distinct aspects of the phenomenon.

An illustrative parallel can be made with the construct of intelligence. Interpreted in a broad sense (as reflected in psychometric batteries of intelligence), working memory is viewed as a facet of the construct of intelligence. However, it would seem plain wrong to use a working memory task to conclude on intelligence as a whole. In the realm of creativity measurement, this level of distinction is often violated: Divergent thinking (DT) as well as product-based measures—tasks leading to a creative production that is then generally rated by external observers—are routinely employed as proxy for "generic" creativity. This is strikingly evident from Snyder, Hammond, Grohman, & Katz-Buonincontro (2019)'s review of 446 studies of creativity spanning three decades, indicating the prevalence of the term *creativity* regardless of the operationalization of the construct. Unfortunately, generalization to the broader construct (creativity) is even less justified than in this context of intelligence research, given that creativity appears even less unitary than does intelligence. This is supported by mounting evidence showing limited domain-generality of creative performance across product-based tasks (e.g., Barbot, 2018a; Barbot, Besançon, & Lubart, 2016), across tasks within the same domain of activity (e.g., Baer, 2011), and even across alternate forms of the same task (Reiter-Palmon, Forthmann, & Barbot, 2019).

In sum, it can fairly be stated that there is no single measure of creativity per se. Across several articles in this collection, a plea for a more accurate identification of the (sub)construct being measured by specific tasks is emphasized (e.g., Acar & Runco, 2019; Barbot, 2019; Kaufman, 2019; Reiter-Palmon, Forthmann, & Barbot, 2019). This point is not trivial: It outlines a need to be more attentive to the use of the creativity vocabulary, by accurately reporting on which "creativity" we are referring to in our studies. We do have the vocabulary and conceptual distinctions to do so, and being more rigorous on that matter would avoid a view of creativity as a loose or elusive entity. A greater accuracy in defining the specific aspects of creativity accounted for in a given study would also increase the accuracy of predictions than can be made about a particular facet of creativity and facilitate metaanalytic work (Barbot, 2018c): Inconsistencies in measurement methods are often the greatest moderator of meta-analytic findings on a range of topics, such as the relationship between creativity and academic achievement (e.g., Gajda, Karwowski, & Beghetto, 2017), self-esteem (Deng & Zhang, 2011), or creativity training program effectiveness (e.g., Scott, Leritz, & Mumford, 2004). Last, but not least, researchers should remain aware of crosscultural differences in the conceptualization of creativity, as well as its relativity vis-a`-vis shifting

definitions, interpretation of task prompts, and even solution space in creative problem-solving tasks (Glaveanu, 2019).

A follow-up question is, then, what are those distinct aspects of creativity, and how to operationalize them? This question calls for a universally accepted taxonomy of measures, which the field is decidedly lacking (Batey, 2012). As noted by Snyder and colleagues (2019), a framework often used as the basis for such a taxonomy is the four Ps model (Rhodes, 1961). This framework has been used to categorize measures according to whether they focus on the *person*, the *process*, the *product*, or the *press* (i.e., environment). It has helped in structuring recent reviews (SaidMetwaly, Van den Noortgate, & Kyndt, 2017) and new taxonomies of measures (Batey, 2012), and it offers a level of specificity that is surely more suitable than having no specificity at all. Reviewing self-report instruments only, Kaufman (2019) organizes measures based on their primary focus: *activities* (i.e., creative participation and achievements), *self-evaluation*, *process*, and *beliefs*. Similarly, Barbot (2019) distinguishes measures of *creative potential, performance*, and *achievement* at the higher level of the taxonomy. This distinction is quite operational and well grounded historically (e.g., Guilford, 1966).

Specifically, measures of *creative potential* (Guilford, 1966) capture the key individual and environmental resources that come into play in the creative work, including aspects of cognition (e.g., DT, convergent and associative thinking), personality (e.g., risktaking, Openness to Experience), emotion, or motivation. Although the distinction between the terms *creative potential* and *creativity* has recently been described as fruitless when it comes to psychological assessment (Silvia, Christensen, & Cotter, 2016), this distinction avoids common confusion as to what creativity is. It places various abilities, motivational factors, and personality trait at the same contributing level (although their individual importance may vary with the task at hand). Accordingly, if DT tasks are often viewed as measures of creativity (Runco & Acar, 2012), it would seem inadequate to qualify a measure of Openness to Experiences as a measure of creativity. In fact, they are both contributing factors of creative activity and as such, can be fairly qualified as measures of creative potential, not creativity.

Indeed, the commonality of *creative potential* measures is that they tap into very specific, somewhat separable resources that come into play in creative work. These resources are numerous, and they all contribute (with often small to moderate individual effect sizes) to potentially multiplicative effects when optimally combined for a given creative work (Sternberg & Lubart, 1991).

> There is much room for different types of creative abilities. What it takes to make the inventor, the writer, the artist, and the composer creative may have some factors in common, but there is much room for variation of pattern of abilities. (Guilford, 1950; p. 451)

What follows, then, is to identify those resources of the creative potential that constitute the common ground for performance in any creative endeavor (Barbot, Tan, Randi,

Santa-Donato, & Grigorenko, 2012). A prominent example is DT, as amply represented in the literature, in this special issue, and later in this article.

Measures of creative performance call upon the whole creative potential as engaged in a given simulated domain-based context ("product-based" tasks), such as producing a drawing, a short story, or a musical composition in standardized condition. Resulting productions are generally rated by external observers using the CAT (Amabile, 1982) or related techniques well represented in this issue (Cseh & Jeffries, 2019; Glaveanu, 2019; Myszkowski & Storme, 2019; Primi, Silvia, Jauk, & Benedek, 2019) and further discussed later. Consistent with decades of findings, these measures are more domain-specific, in that they often engage a different combination of domain-relevant resources for the task at hand, including domain-specific knowledge (Barbot, 2018a).

Finally, measures of *creative achievements* or productivity rely on inventories of past accomplishments in different domains (whether based on self-report or historiometric methodology). Mirroring current practice in the field (Snyder, Hammond, Grohman, & Katz-Buonincontro, 2019), this special issue represents this type of measures only limitedly (Barbot, 2019; Kaufman, 2019). Although these measures could arguably be viewed as the highest standards in creativity assessment due to their objectivity (Simonton, 1999), they also have shortcomings. First, they represent an aspect of the creativity phenomenon that does not apply to many groups (in particular children and adolescents, who are at the forefront of creativity research in education). Second, they may actually not represent a typical sample of behaviors, because creative achievements result from a number of internal and external factors that are not always under the control of the test-takers (e.g., individuals may be lacking the willingness or the opportunity to turn their potential into creative outputs; success of real-life achievements is mainly determined by external factors). Indeed, "purely objective criteria, such as numbers of patents of industrial scientists and engineers, have not served so well probably because they are determined by many irrelevant circumstances" (Guilford, 1966; p. 189). Third, there are cross-cultural differences in the value of creative products, and what constitutes a creative contribution can be drastically different across cultures, contexts, and at different points in time (Glaveanu, 2019).

Main Methods of Creativity Assessment: Challenges and New Developments

In this section, we further discuss the contributions in this special issue that address the main methods for creativity assessment (cf. Forgeard & Kaufman, 2016; Snyder, Hammond, Grohman, & Katz-Buonincontro, 2019): measurement of creative potential with DT and ideation tasks, product-based assessment with a focus on the CAT, and self-report methodology.

Divergences in Divergent Thinking Assessment

Not surprisingly, DT tasks remain a popular choice of methodology in creativity research. In two of the reviews in this special issue (Benedek, Christensen, Fink, & Beaty, 2019; Snyder, Hammond, Grohman, & Katz-Buonincontro, 2019), DT was shown to account for over 50% of the methods across a wide range of creativity studies. Acar and Runco (2019) reiterated the well-known results of validity studies (e.g., Kim, 2006) illustrating that various forms of DT can predict creative behaviors. On those grounds, DT seems to be one of the best ways to quickly obtain information about creative thinking (a central component of creative potential), and there is a growing body of research illustrating that this can be accomplished in person and through online platforms (e.g., Hass, 2015). However, there remain issues regarding their scoring methods, as discussed by Reiter-Palmon and colleagues (2019). They note there are many “researcher degrees of freedom” affecting choices about scoring methods, response exclusion, and data cleaning. Barbot (2019) as well as Reiter-Palmon, Forthmann, & Barbot (2019) also note that there is poor alternate-form reliability for DT, which is mainly a function of the variety of originality scoring methods. That is, whereas originality does not correlate well across different forms of DT tasks (e.g., between alternate-uses performance and consequences performance) or across stimuli used for a given DT task (e.g., between alternate use of a brick or of a cardboard box), fluency is more likely to correlate across different forms. This issue of reliability of DT tasks has long been recognized in the field (Guilford, 1984; Plucker & Runco, 1998), and its impact for both the generalizability of findings and the study of creative thinking change and development is increasingly acknowledged (e.g., Barbot, 2019).

What, then, is divergent thinking, and why does it remain attractive to creativity researchers? First, it represents a basic paradigm to elicit the thinking processes leading to creative responses from participants, which is of central importance in the burgeoning area of neuroscience research on creativity. Benedek and colleagues (2019) emphasize how the field of neuroscience of creativity has adapted the classic DT paradigm into a more trial-based ideation format (e.g., Vartanian et al., 2019), which separates DT processes into ideation time, and response-production time, generally confounded in DT scores. Facilitated by the advent of digital assessments, this alternative format and related extensions are opening new avenues for both neuroscience and behavioral research on creative thinking (Barbot, 2018b). Second, its face validity as an idea generation or brainstorming task cannot be questioned, especially when becreative instructions are used (e.g., Nusbaum, Silvia, & Beaty, 2014). As such, it remains a critical paradigm for testing the effects of interventions designed to augment creative idea generation (e.g., Scott et al., 2004), provided that a number of safeguards are closely followed, as in any study of creativity change and development relying on performance-based tasks (Barbot, 2019).

Indeed, it is clear that DT does not represent a unitary cognitive ability central to the creative process (e.g., Guilford, 1984). If DT were a general cognitive factor, then it could be argued that predictive validity of all forms of DT instruments should be much higher than current estimates suggest, as should alternate-forms reliability. As such, the

recommendations of Reiter-Palmon, Forthmann, & Barbot (2019)—that several forms of DT tasks be used in general DT studies—should be adhered to if one is interested in drawing broad conclusions about this fundamental dimension of creative potential in both experimental and individual differences research. Although a comprehensive assessment of DT would currently require an unreasonable number of tasks to account for all its possible products and content facets (Guilford, 1984), relying on a range of tasks and clearly identifying the nature of these DT tasks (e.g., using Guilford's, 1984, structure of intellectbased taxonomy) seem to be a reasonable way to achieve greater coherence in this line of work (Reiter-Palmon, Forthmann, & Barbot, 2019). This recommendation resonates well with Snyder, Hammond, Grohman, & Katz-Buonincontro's (2019) advice regarding the need to distinguish domain and task diversity in creativity as well as relying on multimethod and mixed methods.

As always, the metrics for obtaining information from DT responses beyond fluency are those that will provide the most amount of information vis-a`-vis creative potential. The articles in this special issue illustrate that there is still no consensus on the best way to obtain that information. It follows that not only should scoring match instructions (Reiter-Palmon, Forthmann, & Barbot, 2019) but the choice of scoring should also match the theoretical aims of the research project, which may be quite different from study to study and from culture to culture (Glaveanu, 2019). Acar and Runco (2019) argue that semantic scoring is the way forward in objectively measuring remoteness in DT responding, but that argument is incomplete. Indeed, latent semantic analysis (LSA) shows promise for evaluating cognitive mechanisms underpinning some manifestations of DT (i.e., those relying on semantic classes), but due to the nature of LSA itself, this approach may not be sufficient for directly assessing creativity or originality (cf. Acar & Runco, 2014; Forthmann, Oyebade, Ojo, Günther, & Holling, 2018; Hass, 2017). The most obvious illustration of this point is that LSA may prove unusable in scoring figural DT tasks (e.g., repeated-figures or incomplete-figures tasks), whereas much DT production (and creative thinking in general) is indeed not necessarily processing semantic content (e.g., visual, auditory, symbolic).

Substantial progress is also being made with regard to scoring using the well-worn uniqueness method tied to the statistical frequency of responses. Olteteanu, Hass, and Zunjani (2018) initiated the construction of a set of normative responses for 420 objects that can be used as prompts in the alternate-uses task. They used a computer algorithm for clustering similar responses that has the potential to limit the researcher's degrees of freedom involved in grouping similar responses together, for both flexibility and various kinds of originality scoring. These kinds of norming efforts have potential to support a greater consensus, at least at the prescoring stage (i.e., categorization of response, determination of their adequacy) and may surely provide more reliable flexibility scores across different studies.

Product-Based Assessment: How to Skin a CAT?

Since Amabile (1982) proposed the CAT almost four decades ago, the use of this approach as a way to evaluate and measure creative performance in a wide variety of product-based tasks has increased, making it one of the most common ways to assess it. This is evident from the solid upward trend in the use of productbased assessment suggested by Snyder, Hammond, Grohman, & Katz-Buonincontro's (2019) review. Beyond the range of tasks that could lead people to generate creative outputs, the core of product-based assessment lies in the way they are then scored for their creative quality. This approach was already common in creativity research prior to Amabile's contribution, but the formalization of the CAT has further established its legitimacy.

Several articles in the special issue directly discuss the CAT approach (Cseh and Jeffries, 2019; Glaveanu, 2019; Myszkowski & Storme, 2019). Cseh and Jeffries (2019) cover what we know about the CAT from these last 35 years of research. More important, their article focuses on the inconsistent ways in which this approach has been used and operationalized in practice. They discuss the potential effects of these different operationalizations on the integrity of results, and they suggest that these may be quite impactful. Given that judges or raters are critical to this approach, understanding who they are and what their level of expertise is, is also important. Further, Cseh and Jeffries also address issues related to rating scales used, how productions are presented, and variations in the evaluation of interrater reliability. Following on this latter issue, Myszkowski and Storme (2019) challenge the classic test theory paradigm in aggregating information from raters and suggest an alternative way to evaluate the CAT's interrater reliability that is based on item response theory (IRT) under a framework that they coined judge response theory. They argue that this approach is a more appropriate and accurate way to evaluate interrater reliability, notably because it can attach a distinct level of reliability to productions of a distinct level of rated creativity. Indeed, much of the creativity research employing that CAT reports Cronbach's alpha as an index of interrater reliability (Cseh & Jeffries, 2019). However, there are longstanding disagreements as to whether this is the appropriate index for assessing interrater reliability. More important, Myszkowski and Storme (2019) emphasize how IRT-based scoring can lead to a more accurate estimation of the latent trait (i.e., the creative "value" of the rated productions), which greatly questions common practices regarding the aggregation of ratings.

Supported by the empirical demonstration of Primi, Silvia, Jauk, & Benedek (2019), Myszkowski and Storme (2019) also draw an appealing picture of what product-based assessment could look like in future creative performance research. This includes the parsimonious selection of complementary (rather than redundant) raters, which can be different according to the information already available for a given production. Both Myszkowski and Storme's proof of concept and Primi et al.'s empirical illustrations are demonstrating the flexibility of the IRT frameworks to actually accommodate most of the variations in the use of the CAT outlined by Cseh and Jeffries (2019). Indeed, one can easily imagine larger scale studies relying on product-based tasks that make use of raters of various levels of expertise, rating different productions

(both in quantity and quality), and using rating scales that are not necessarily uniform. This flexibility could fundamentally reshape the way we use the CAT. Further, understanding “differences in individual judgment as a psychological process” (Myszkowski & Storme, 2019) appears particularly appealing to address some of the challenges outlined by Glaveanu (2019) regarding intercultural CAT research.

Together, what is clear from these articles is that although many researchers use what they term CAT, the application of this approach is not universal. There are important questions regarding the use and application of the CAT, which will impact not only the individual choices of researchers but also of the entire creativity research community. Although the prospects offered by Myszkowski and Storme (2019) as well as Primi, Silvia, Jauk, & Benedek (2019) are appealing, it is also clear that more studies and new approaches are needed to directly evaluate the utility and feasibility of the CAT under certain conditions. The research community has started to address these issues (see e.g., work on expert vs. nonexpert raters), but more research is needed to confirm that these new approaches are worth the effort with respect to the validity of creative performance scores obtained, as well as with regard to potential issues of implementation (e.g., statistical packages available to extract IRT-based scores; logistics of complex rating designs).

Bringing Back the Self in Self-Report Methodologies

Self-report methodology has received surprisingly positive praise in this special issue (e.g., Barbot, 2019; Cotter & Silvia, 2019; Kaufman, 2019; Karwowski, Han, & Beghetto, 2019) and remains a popular way to assess aspects of creativity as identified by Snyder, Hammond, Grohman, & Katz-Buonincontro (2019) and closely reviewed by Kaufman (2019)—in particular, variables of *potential* and *achievement activities*). This contrasts with typical criticisms from the field of creativity (e.g., “All self-report measures are somewhat suspect and perhaps especially so in the area of creativity” Baer, 2011; p. 310). The general shortcoming of self-report instruments may be mainly attributed to how researchers make inferences from this methodology (Spector, 1994). If we intend to measure “generic creativity” using the sole reliance on test-takers’ perceptions of their own creative abilities through self-report measures, then we are surely missing the point: It is established that people are usually inaccurate in assessing their own creativity (i.e., self-rated creativity; Reiter-Palmon, RobinsonMorral, Kaufman, & Santo, 2012). However, despite evident limitations of this methodology in creativity research (for a review, see McKibben & Silvia, 2017), there are some constructs relevant to creativity that are thus far best measured with self-report methodology because they precisely tap into people’s own subjectivity and perceptions (Kaufman, 2019).

Of course, self-report methodology has been the preferred approach to capture aspects of creative personality and motivation, and it is probably the only way that people can report on their day-to-day creative accomplishment and/or participation in creative endeavors (Cotter & Silvia, 2019; Kaufman, 2019). But there is also a range of creative self-related constructs (e.g., creative self-efficacy, creative mind-sets, creative

identity) that have strongly (re)surfaced in the field in these past few years (Karwowski & Kaufman, 2017) and certainly contribute to the greater enthusiasm toward this methodology. These creative self-related variables, uniquely operationalized using self-report methodology, represent relevant aspects of the creative potential because they predict real-life creative performance or achievements (e.g., Royston & ReiterPalmon, 2017) to the same or even a greater extent than DT tasks do.

They also tap into unique aspects of the potential that other measures of creative potential and performance (e.g., product-base assessment) cannot capture, such as one's own willingness and agency in creative pursuits. This point is important because if the goal of creativity research is to bolster the fulfilment of creative potential (e.g., Runco, 2016), then we need to better understand what makes people turn their potential into action and invest creative pursuits. People may have high creative potential under the perspective of a DT task performance and other indicators of potential, but will this potential translate into action, that is, into real-life creativity? This particular question is rarely addressed in creativity research, and the (re)emergence of creative self-related studies is a unique opportunity to fill this gap. This is illustrated by Karwowski, Han, & Beghetto (2019), showing how performance on classic creative performance– based tasks is intricately and dynamically related to people's self-confidence in their creative abilities.

This focus on creativity "dynamics" also resonates with Barbot's (2019) methodological recommendations on the study of creativity change and development, outlining that, contrary to many other methodologies, self-report measures may better fit the constraints of repeated-measurement designs (although they are not exempt from their own sets of issues). Pushed to the extreme, experience sampling methods, for which important guidelines and methodological recommendations have been outlined by Cotter and Silvia (2019), rely quite heavily on repeated administration of self-report items, illustrating how this methodology can help address unique questions regarding the dynamics of creativity in more ecologically valid contexts.

In all, there are good reasons to be enthusiastic about a better valuation of self-report methodology in creativity research, as long as self-report measures are taken for what they are (e.g., indicators of creative self-efficacy, creative mind-set) and not measures of creativity per se.

Conclusions: Toward New Standards in Creativity Assessment

The collection of articles in this special issue has touched upon the most critical challenges in our field regarding creativity assessment. Together, a number of principles and recommendations for best practices in creativity assessment can be derived. This is a difficult exercise because the contributions in this special issue do not converge to definitive conclusions. However, there seems to be a general consensus on the need to adhere to some basic standards, before the field can agree on any gold standard.

Those basic standards can be summarized under the terms *transparency* and *homogenization* of creativity assessment.

Regarding *transparency* of creativity assessment, the consideration is mainly on the accuracy of defining, operationalizing, and adequately reporting on what any score permits one to validly conclude. If there is an acceptable consensus on a general definition of *creativity*, as noted throughout the special issue, virtually none of the existing measurement approaches are a direct operationalization of that general definition. In short, there is no such thing as "all-purpose creativity tests" (Baer, 2011, p. 312), or generic-creativity measures (Barbot, 2018c). This is not necessarily a challenge, as long as the particular facet(s) of the creativity phenomenon being investigated are themselves clearly defined and operationalized. It goes without saying that the selected facet(s) and corresponding operationalization should match the research objective and that resulting study findings should be accurately reported in light of this operationalization. In practice, a study employing for example an alternate-uses task taps into divergent production of semantic classes applied to unusual uses for common objects, one of numerous facets and operationalization of DT (Guilford, 1984), itself a key cognitive resource of the creative potential. Of course, it is tempting to use a shortcut and label it a *creativity task*, but doing so would fail this principle of transparency in operationalization and reporting. On a related note, selfreport measures are not necessarily inferior to performance-based tasks such as DT at operationalizing creativity: In a way, they are all suboptimal, in that they incompletely represent the broader phenomenon. The point is that they are simply addressing different, but potentially equally important, aspects of the phenomenon. Either way, relying on a single method to measure any aspect of creativity is surely a perilous venture.

Regarding *homogenization* of creativity assessment, several contributions in this special issue have emphasized how much variability there is in the actual implementation of measures across studies and settings (e.g., variations in DT tasks, CAT ratings, neuroscience paradigms). For the same operationalization of a given facet of creativity (e.g., DT as a key component of creative potential), a desirable attainment for our field is the homogenization of measures–tasks and their administration conditions used in creativity research (e.g., instructions, time on task). This point corresponds to Rossiter's (2018) call for the identification of an "*optimal standard measure* of each major construct." If this endeavor is unrealistic for the general term *creativity*, it is reasonably achievable at the subconstruct or facet level. Reasonably achievable does not mean that it will be effortless. It is probably our biggest challenge for the years to come.

Specifically, various measures of specific facets of creativity may be highly standardized in a given study, but even modest variations across studies often shift the substantive meaning of the scores obtained. This is illustrated by mounting evidence from the DT research line (Reiter-Palmon, Forthmann, & Barbot, 2019) showing variations in performance relative to instructions, time on task, or stimulus used. These

differences inform to some extent the nature of the underlying facet of creativity being measured. Work that precisely investigates the optimal conditions to elicit this facet (e.g., are DT tasks better under a 2-min, 10-min, or self-paced format?) are critical for the field if their ultimate outcome is the formalization of a definite task or procedure format that can be used broadly.

These efforts are necessary to reach, more generally, the research standards that the field of psychology is seeking to meet, in particular with respect to replication (cf. PACA' special section "Replications in Psychology," Plucker, 2014). Follow-up questions regarding creativity assessment issues include whether the creativity research community can reach an acceptable consensus on the optimal assessment methods, how they will be determined, and which mechanisms can be used to establish a homogeneous use of these methods. Determining the best format and methods should undoubtedly be evidencedbased, but creativity researchers will still have to "be creative" to settle on the answers to the other questions. Our era of open-science, participatory mechanisms, and online assessment provides all the tools we need to achieve these goals. Initiatives such as the International Personality Item Pool (e.g., Goldberg et al., 2006) or the Cognitive Atlas (Poldrack et al., 2011) have tackled similar issues that we are currently facing. These initiatives could fundamentally inspire future development in creativity assessment, and it is our hope that this special issue will stimulate research efforts and discussion in this direction.

References

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. Creativity Research Journal, 26, 229 –238. http://dx.doi.org/10.1080/10400419.2014.901095

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

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology, 43, 997–1013. http://dx.doi.org/10.1037/0022-3514.43.5.997

Baer, J. (2011). How divergent thinking tests mislead us: Are the Torrance Tests still relevant in the 21st century? The Division 10 debate. Psychology of Aesthetics, Creativity, and the Arts, 5, 309 –313. http://dx.doi.org/10.1037/a0025210

Barbot, B. (2018a). Creativity and self-esteem in adolescence: A study of their domain-specific, multivariate relationships. Journal of Creative Behavior. Advance online publication. http://dx.doi.org/10.1002/jocb.365

Barbot, B. (2018b). 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. (2018c). “Generic” creativity as a predictor or outcome of identity development? Creativity: Theories-Research-Applications, 5, 159 –164. http://dx.doi.org/10.1515/ctra-2018-0013

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

Barbot, B., Tan, M., Randi, J., Santa-Donato, G., & Grigorenko, E. L. (2012). Essential skills for creative writing: Integrating multiple domainspecific perspectives. Thinking Skills and Creativity, 7, 209 –223. http://dx.doi.org/10.1016/j.tsc.2012.04.006

Batey, M. (2012). The measurement of creativity: From definitional consensus to the introduction of a new heuristic framework. Creativity Research Journal, 24, 55–65. http://dx.doi.org/10.1080/10400419.2012.649181

Benedek, M., Christensen, A. P., Fink, A., & Beaty, R. E. (2019). Creativity assessment in neuroscience research. Psychology of Aesthetics, Creativity, and the Arts, 13, 218 –226. http://dx.doi.org/10.1037/aca0000215

Binet, A. (1900). L’observateur et l’imaginatif. [The observer and the imaginative]. L’Année Psychologique, 7, 519 –523. http://dx.doi.org/10.3406/psy.1900.4682

Cotter, K. N., & Silvia, P. J. (2019). Ecological assessment in research on aesthetics, creativity and the arts: Basic concepts, common questions, and gentle warnings. Psychology of Aesthetics, Creativity, and the Arts, 13, 211–217. http://dx.doi.org/10.1037/aca0000218

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

Deng, X.-P., & Zhang, X.-K. (2011). Understanding the relationship between self-esteem and creativity: A meta-analysis. Xinli Kexue Jinzhan, 19, 645– 651.

Forgeard, M. J., & Kaufman, J. C. (2016). Who cares about imagination, creativity, and innovation, and why? A review. Psychology of Aesthetics, Creativity, and the Arts, 10, 250 –269. http://dx.doi.org/10.1037/aca0000042

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2018). Application of latent semantic analysis to divergent thinking is biased by elaboration. Journal of Creative Behavior. Advance online publication. http://dx.doi.org/10.1002/jocb.240

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

Fürst, G. (2018). Measuring creativity with planned missing data. Journal of Creative Behavior. Advance online publication http://dx.doi.org/10.1002/jocb.352

Gajda, A., Karwowski, M., & Beghetto, R. A. (2017). Creativity and academic achievement: A meta-analysis. Journal of Educational Psychology, 109, 269 – 299. http://dx.doi.org/10.1037/edu0000133

Galton, F. (1869). Hereditary genius: An inquiry into its laws and consequences (Vol. 27). http://dx.doi.org/10.1037/13474-000

Glaveanu, V. P. (2019). Measuring creativity across cultures: Epistemological, methodological and ethical considerations. Psychology of Aesthetics, Creativity, and the Arts, 13, 227–232. http://dx.doi.org/10.1037/aca0000216

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84 –96. http://dx.doi.org/10.1016/j.jrp.2005.08.007

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

Hass, R. W. (2015). Feasibility of online divergent thinking assessment. Computers in Human Behavior, 46, 85–93. http://dx.doi.org/10.1016/j.chb.2014.12.056

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. Memory & Cognition, 45, 233–244. http://dx.doi.org/10.3758/s13421-016-0659-y

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

Jankowska, D. M., Czerwonka, M., Lebuda, I., & Karwowski, M. (2018). Exploring the creative process: Integrating psychometric and eyetracking approaches. Frontiers in Psychology, 9, 1931. http://dx.doi.org/10.3389/fpsyg.2018.01931

Karwowski, M., Han, M.-H., & Beghetto, R. A. (2019). Toward dynamizing the measurement of creative confidence beliefs. Psychology of Aesthetics, Creativity, and the Arts, 13, 193–202. http://dx.doi.org/10.1037/aca0000229

Karwowski, M., & Kaufman, J. C. (2017). The creative self. Retrieved from https://www.elsevier.com/books/the-creative-self/karwowski/978-0-12-809790-8

Kaufman, J. C. (2019). Self-assessments of creativity: Not ideal, but better than you think. Psychology of Aesthetics, Creativity, and the Arts, 13, 187–192. http://dx.doi.org/10.1037/aca0000217

Kim, K. H. (2006). Can we trust creativity tests? A review of the orrance Tests of Creative Thinking (TTCT). Creativity Research Journal, 18, 3–14. http://dx.doi.org/10.1207/s15326934crj1801_2

Loesche, F., Goslin, J., & Bugmann, G. (2018). Paving the way to Eurekaintroducing “Dira” as an experimental paradigm to observe the process of creative problem solving. Frontiers in Psychology, 9, 1773. http://dx.doi.org/10.3389/fpsyg.2018.01773

McKibben, W. B., & Silvia, P. J. (2017). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. Journal of Creative Behavior, 51, 57– 69. http://dx.doi.org/10.1002/jocb.86

Myszkowski, N., & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgements. Psychology of Aesthetics, Creativity, and the Arts, 13, 167–175. http://dx.doi.org/10.1037/aca0000225

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.

Olteteanu, A., Hass, R. W., Zunjani, F. (2018). Computational extraction of metrics and normative data on the alternative uses test on a set of 420 household objects. Manuscript under review.

Partnership for 21st Century Skills. (2008). 21st century skills, education & competitiveness: A resource and policy guide. Tucson, AZ: Author.

Plucker, J. A. (Ed.). (2014). Replications in psychology [special section]. Psychology of Aesthetics, Creativity, and the Arts, 8, 2–29. Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), The Cambridge handbook of creativity (pp. 48 –73). http://dx.doi.org/10.1017/CBO9780511763205.005

Plucker, J. A., & Runco, M. A. (1998). The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. Roeper Review, 21, 36 –39. http://dx.doi.org/10.1080/02783199809553924

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y.,... Bilder, R. M. (2011). The Cognitive Atlas: Toward a knowledge foundation for cognitive neuroscience. Frontiers in Neuroinformatics, 5, 17. http://dx.doi.org/10.3389/fninf.2011.00017

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying manyfacet rasch modeling in the assessment of creativity. Psychology of Aesthetics, Creativity, and the Arts, 13, 176 –186. http://dx.doi.org/10.1037/aca0000230

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. Psychology of Aesthetics, Creativity, and the Arts, 13, 144 –152. http://dx.doi.org/10.1037/aca0000227

Reiter-Palmon, R., Robinson-Morral, E. J., Kaufman, J. C., & Santo, J. B. (2012). Evaluation of self-perceptions of creativity: Is it a useful criterion? Creativity Research Journal, 24, 107–114. http://dx.doi.org/10.1080/10400419.2012.676980

Rhodes, M. (1961). An analysis of creativity. Phi Delta Kappan, 42, 305–310. Rossiter, J. R. (2018). The new psychometrics: Comment on Appelbaum et al. (2018). American Psychologist, 73, 930 –931. http://dx.doi.org/10.1037/amp0000342

Royston, R., & Reiter-Palmon, R. (2017). Creative self-efficacy as mediator between creative mindsets and creative problem-solving. Journal of Creative Behavior. Advance online publication. http://dx.doi.org/10.1002/jocb.226

Runco, M. A. (2016). Commentary: Overview of developmental perspectives on creativity and the realization of potential. New Directions for Child and Adolescent Development, 2016, 97–109. http://dx.doi.org/10.1002/cad.20145

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. Creativity Research Journal, 24, 66 –75. http://dx.doi.org/10.1080/10400419.2012.652929

Said-Metwaly, S., Van den Noortgate, W., & Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. Creativity: Theories-Research-Applications, 4, 238 –275. http://dx.doi.org/10.1515/ctra-2017-0013

Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. Creativity Research Journal, 16, 361–388. http://dx.doi.org/10.1080/10400410409534549

Silvia, P. J., Christensen, A. P., & Cotter, K. N. (2016). Commentary: The development of creativity—Ability, motivation, and potential. New Directions for Child and Adolescent Development, 2016, 111–119. http://dx.doi.org/10.1002/cad.20147

Simonton, D. K. (1999). Creativity from a historiometric perspective. In R. J. Sternberg (Ed.), Handbook of creativity (pp. 116 –133). New York, NY: Cambridge University Press.

Snyder, H., Hammond, J. A., Grohman, M. G., & Katz-Buonincontro, J. (2019). Creativity measurement in undergraduate students from 1984 – 2013: A systematic review. Psychology of Aesthetics, Creativity, and the Arts, 13, 133–143. http://dx.doi.org/10.1037/aca0000228

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. Journal of Organizational Behavior, 15, 385–392. http://dx.doi.org/10.1002/job.4030150503

Sternberg, R. J. (2018). What's wrong with creativity testing? Journal of Creative Behavior. Advance online publication. http://dx.doi.org/10.1002/jocb.237

Sternberg, R. J., & Lubart, T. I. (1991). An investment theory of creativity and its development. Human Development, 34, 1–31. http://dx.doi.org/10.1159/000277029

Torrance, E. P. (1998). The Torrance Tests of Creative Thinking–Norms— Technical manual research edition—Verbal tests, Forms A and B—Figural tests, Forms A and B. Princeton, NJ: Personnel Press.

Vartanian, O., Beatty, E. L., Smith, I., Forbes, S., Rice, E., & Crocker, J. (2019). Measurement matters: The relationship between methods of scoring the alternate uses task and brain activation. Current Opinion in Behavioral Sciences, 27, 109 –115. http://dx.doi.org/10.1016/j.cobeha.2018.10.012