The Black Box of the Consensual Assessment Technique: Some Questions and Doubts on the Subjective Rating of Creativity

Xavier Caroff and Justine Massu

Keywords Consensual Assessment Technique · Creativity · Methodology

From the simple idea to a concrete realization, to what extent can we consider that a production in any particular domain is really creative? How can we evaluate that this production is both original and adapted to its context?

These questions can be answered from two different approaches. The first consists of using standardized rating scales in order to evaluate the creativity of a production. Many rating scales can be found and the most used is certainly the "Creative Product Semantic Scale" (CPSS; O'Quin & Besemer, 1989). In this case, the evaluation of a creative production is based on three different dimensions: Novelty, Resolution, and Elaboration and synthesis. This scale presents quite acceptable metric qualities (for example, Besemer, 1998, 2000; Besemer & O'Quin, 1986, 1999; O'Quin & Besemer, 1989, 2006). However, the use of such a scale has been largely discussed because it presents the disadvantage of relying on a particular theoretical conception of creativity that is the conception of the authors of the scale. Therefore the proposed rating criteria appear to be scarcely objectively specified (for example, Amabile, 1996; Kaufman et al., 2008).

The second approach is based on subjective ratings collected from people suited and competent to estimate creativity. Amabile (1996), then Hennessey et al. (2011) proposed a brief history of these methods which seems to originate in Galton's work on eminence. The same authors also noted various objections toward these methods. First, subjective evaluation seems relatively disconnected from scientific conceptions

X. Caroff (✉) · J. Massu
Université Paris Cité and Univ. Gustave Eiffel, LaPEA, Boulogne-Billancourt F-92100, France
e-mail: xavier.caroff@u-paris.fr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

T. Lubart et al. (eds.), *Homo Creativus*, Creativity in the Twenty First Century,
https://doi.org/10.1007/978-3-030-99674-1_11

of creativity. Second, the typical application of this approach does not allow for the distinction between the evaluation of creativity and proximal characteristics such as esthetic or technical qualities.

Therefore, Amabile developed a method called the "Consensual Assessment Technique" (CAT) (Amabile, 1982, 1996; Hennessey et al., 2011). Since the first publication, the essential of the CAT methodological principles did not evolve. Actually, the CAT is used frequently in research on creativity, making it the "Gold Standard" of creativity evaluation (Carson, 2006). Kaufman et al. (2008) identified at least three reasons to explain such popularity among researchers: the CAT rates creativity such as it can be observed through simple productions, it does not rely on a particular theoretical conception of creativity and it fits how creativity is evaluated concretely in everyday life.

1 The Consensual Assessment Technique

Originally, the CAT has been conceived from a clear distinction between two definitions of creativity. The first definition corresponds to researchers' conception of creativity whereas the second is more operational and is based on the implicit conceptions of individuals requested to evaluate the creativity of a production. According to Amabile (1996, p. 35), « a product or response will be judged as creative to the extent that (a) it is both a novel and appropriate, useful, correct or valuable response to the task at hand, and (b) the task is heuristic rather that algorithmic». This conception is steeped into the standard definition of creativity which has a long history (e.g., Runco & Jaeger, 2012) and has been progressively adopted by most of the researchers in the field. According to Amabile (1996), criteria that enable the identification of creative productions can neither be defined nor objectively measured (see also, Runco & Jaeger, 2012). For this reason, it is necessary to rely on subjective criteria. Then, she proposed an operational definition called the "consensual definition of creativity" and stated that: "a product or response is creative to the extent that appropriate observers independently agree it is creative. Appropriate observers are those who are familiar with the domain in which the product was created or the response articulated. Thus, creativity can be regarded as the quality of products or responses judged to be creative by appropriate observers, and it can also be regarded as the process by which something so judged is produce" (Amabile, 1996, p. 33). Thus, the consensual definition identified creativity through the process of evaluation. However, these definitions fill different functions while being closely linked: "In essence, the conceptual definition is a best guess as to what characteristics appropriate observers are looking for when they assign ratings of 'creativity' to products" (Amabile, 1996, p. 37).

1.1 Methodological Principles

Therefore, the CAT represents the operationalization of the consensual definition of creativity. However, there are methodological principles that need to be respected by researchers in order to make optimal use of the CAT (for a critical review see Cseh & Jeffries, 2019). These principles refer first to the characteristics of the productions under evaluation, and second to the evaluation procedure. Because these principles have been well documented previously (e.g., Amabile, 1982, 1996; Baer & McKool, 2009; Hennessey et al., 2011; Kaufman et al., 2008), we will briefly present them.

Regarding the first principle, the productions under evaluation should have been created under open-ended work conditions. In this way, there is a greater chance to obtain enough variability and novelty in the answers of different subjects. Moreover, whatever the domain of production is, the work conditions need to be accessible so judges can easily rate them. Second, regarding the evaluation procedure, we need to ensure that the judges have sufficient experience in the domain of endeavor in order to be able to identify productions that are creative. Regarding the assessment process, productions should be presented in a random order. Then judges should rate independently the level of creativity for each production relative to the others and in accordance with their own conception of creativity (Baer & Kaufman, 2019). Hence, researchers should not provide empirical criteria or a definition of creativity. Moreover, it is recommended to ask judges to rate alternative dimensions besides creativity such as esthetic or technical qualities. In this way, it is possible to appreciate the extent to which the rating of creativity has been made independently from other related characteristics of the productions.

1.2 Statistical Validity

Different strategies of statistical analysis can be used in order to establish the validity of the creativity ratings obtained via the CAT. Discriminant validity enables us to verify the extent to which the creativity ratings are independent from alternative dimensions (Amabile, 1996). Concretely, it can be tested using two different strategies. The first consists of asking judges to rate the productions on different criteria and then to perform a factor analysis in order to test if one factor gathers the creativity relevant criteria and if this factor can be isolated from others gathering the alternative dimensions. In a study conducted by Amabile (1996, study 1), collages produced by children were rated on several criteria. Results from the analysis distinguished two relatively independent factors, creativity and technical goodness. These factors have also been found by Čorko and Vranić (2004) in a study using the same procedure. In a second study conducted by Amabile (1996, study 14), judges were asked to rate creativity among other criteria of poems written by students. In contrast, the analysis discerned in this case three different factors (creativity, style and technique) and the creativity factor was less clearly distinct than in previous studies.

The second strategy to assess discriminant validity consists of correlating a score of creativity with the rating of another dimension. A first series of results focused on the correlation between creativity and technical goodness (Amabile, 1996, study 1 to 8). But we notice that the correlation actually depends on respective reliabilities of the two rated dimensions. When reliabilities are between 0.70 and 0.80, the correlations between creativity and technical goodness vary from 0.13 to 0.28. However, when reliabilities are between 0.80 and 0.90, correlations vary from 0.70 to 0.77. Čorko and Vranić (2004) obtained a correlation of 0.63 between these two dimensions, considering that the weakest interrater agreement was 0.77. Thus, if we take into account this dampening effect, we should conclude that the ratings of creativity are quite considerably correlated with the technical goodness of productions.

However, most researchers do not test the discriminant validity of ratings obtained from judges. They are satisfied with the reliability analysis of creativity ratings. It is effectively important to test reliability in order to ensure that the variance of creativity ratings is explained by differences in the estimated creativity level rather than the error variance. Since the first research conducted by Amabile (1982, 1996; Hennessey et al., 1999), most of CAT based research considered that the interrater agreement regarding creativity ratings of productions was quite acceptable. This point is essential because the validity of the subjective evaluation, upon which the CAT is based, relies on the interrater reliability. Indeed, unlike the classical psychometric approach that seeks to test the validity of a measure, in this case the validity results from a reasoned argumentation stating that *if* the selected judges to rate creativity are familiar with the domain of production, and *if* the interrater reliability is high, *then* what raters agree on is necessarily creativity.

The following sections of this chapter will precisely question the principle of this line of argument. In the first part we will examine what types of judges should be selected in order to rate the creativity of productions from specific domains. We will also discuss how the judges' characteristics can influence their ratings. In the second part, we will try to identify the content of the implicit conceptions on creativity that guides the subjective evaluation of productions. Subsequently, we will verify if these implicit conceptions correspond to scientific theories and we will seek to understand the extent to which judges from the same domain of expertise have consistent implicit conceptions of creativity.

2 What Types of Judges Should be Selected to Rate the Creativity of Productions?

As stated by Amabile (1982, 1996), the validity of the consensual assessment relies on the selection of the individuals who will be asked to rate the creativity of productions. However, this author's stance seems to have evolved since. Indeed, in a first version of her reference work, she suggested that judges should be experts of the domain. According to Baer and Kaufman (2019), expertise is a necessary condition to ensure

the CAT validity. However, they do not provide evidence for such necessity nor define clearly what constitutes expertise. In most research using the CAT, expertise has been left behind and replaced by a simpler criterion of familiarity with the domain. Similarly, we can observe that the author's position evolved regarding the judges sample composition. Initially, Amabile considered that "the level of experience for all judges need not to be identical" (Amabile, 1996, p. 41). Then, she simplified the criterion, considering it sufficient if « they have roughly equivalent experience with the domain in question» (Hennessey et al., 2011, p. 255).

However, the CAT is based on the premise that if judges are sufficiently familiar with the domain of production they should be able to rate spontaneously the creativity of these productions. Following this logic, it is unnecessary, if not detrimental for Amabile (1982, 1996; Hennessey et al., 2011) to give the experts a normative definition of creativity, to ask them to rate productions according to explicit criteria (such as the different aspects from the theoretical definition: originality and appropriateness), or even to train them to use such criteria. Consequently, we make the assumption that judges who are sufficiently familiar with a certain domain should agree on the extent to which the productions of this domain are creative. These premises have been shared tacitly and most often explicitly by every researcher using the CAT (for example, Baer & McKool, 2009; Kaufman & Baer, 2012; Kaufman et al., 2008).

3 Which Criteria Should We Use to Select Judges to Rate Creativity?

Selecting a group of appropriate judges should depend on two criteria. The domain from which the rated production belongs but also the type of objective pursued through this evaluation: practical or scientific. It is in this sense that the term appropriate should be understood and applied when selecting judges.

However, regarding the first criteria, the notion of familiarity with the domain is insufficient to select appropriate raters according to Kaufman and Baer (2012). They noted that it is theoretically possible to observe a high degree of consensus in a group of raters who are relatively novice within the domain of production. Thus, they consider that it is crucial to seek a high degree of interrater agreement from a truly expert group from the specific domain of production. Research has recently examined different levels of expertise and their effects on the reliability of creativity ratings.

Following these principles, Kaufman and Baer (2012) identified three types of judges according to their level of expertise: (1) the experts who have at least ten years of specific experience within the domain and have received an honor for their exceptional realizations within the same domain. (2) quasi experts that are experienced but have not been recognized for their expertise, and (3) novices that have no expertise in the domain but have skills that are related to the type of production (such

as graduate students, teachers or professors on creativity). According to the same authors, if this distinction between the types of judges is relevant, we should observe differences in results among the groups. Within the experts' group, the interrater agreement should be high, much higher than within the group of quasi experts and or even the group of novices. Next, between the clusters, experts' rating should have a weak correlation, if any at all, with quasi experts' ratings and even less with novices' ratings.

Such predictions have been tested by two kinds of research in creativity. The first kind consisted of comparing research that selected only experts or quasi experts to rate creative productions to research that opted for novices as judges. The scientific objectives of the selected research were not taken into account. The aim of this comparison was too investigate if the reliabilities of the ratings were different between the different categories of judges. We will present the principal results regarding the comparison between experts and novices. Readers who are interested in more details on the comparison between experts and quasi experts can refer to a synthesis published by Kaufman and Baer (2012).

Following the first research published by Amabile (1982, 1996), it has been demonstrated that experts' ratings of creativity had good reliability (for example, Baer, 1997, 2003; Baer et al., 2004). To our knowledge, few studies failed to demonstrate that experts had an acceptable interrater reliability (Gerrard et al., 1996; Hickey, 2001). However, this rarity might only be due to the unlikeliness of finding published research showing that expert judges had insufficient interrater reliability. Regarding novices rating creativity, we can observe that interrater reliability can reach and even exceed the conventional cutoff of 0.70–0.80. This cutoff will be discussed later in this chapter. This high reliability was found in research on artistic or literary creativity where selected judges were students enrolled in an artistic program, which accredits a certain familiarity with the domain, but also students enrolled in a non-related program (for example, Baer, 1996; Chen et al., 2002, 2005; Joussemet & Koestner, 1999; Kasof et al., 2007; Niu & Sternberg, 2001). According to Kaufman and Baer (2012), if these results confirm that novices' ratings can result in a consensual evaluation, this does not indicate the validity of the evaluation because the raters are not experts.

A second kind of research tests more precisely the predictions made by Kaufman and Baer (2012). In this line of research, ratings of judges from different levels of expertise are compared but in rigorously equivalent conditions. In a first publication on the CAT, Amabile (1982; experiment 1) asked children from seven to eleven years old, with limited creative abilities, to make collages using pre-cut pieces of paper. The creativity of these productions were rated by three types of judges with different levels of expertise: members of the Stanford University psychology department (faculty and graduate students), elementary- and secondary-school art teachers (who happened to be taking a course at Stanford), undergraduate and graduate artists from the art department at Stanford University, each of whom had spent at least 5 years working in studio art. Results show that psychologists' ratings had a relatively acceptable consistency ($\alpha = 0.73$), but weaker than art teachers' ratings ($\alpha = 0.88$) and almost equivalent to artists' ratings ($\alpha = 0.77$). Furthermore, the correlation between psychologists and art teachers' ratings was too weak ($r = 0.44$)

to conclude that these judges agree on their evaluation of creativity productions. But this correlation was slightly higher between art teachers and artists' ratings ($r = 0.65$). These results are particularly interesting because they illustrate the difficulties we can encounter in research comparing the evaluations of various types of experts. These difficulties relate both in the selection of appropriate judges and in the interpretation of the results. Kaufman and Baer (2012) presented this work in their literature review as an example of comparative research on experts and quasi experts' ratings. They consider that «although the psychologists lacked artistic expertise, they did have a different type of expert knowledge (i.e., understanding children) that might have been relevant to making these judgments, and thus cannot be considered complete novices» (p. 87). They pointed out also that according to Amabile (1996), appropriate judges should have at least a certain level of educational background and experience in the specific domain of production. If we agree with these statements, then who can be considered as a complete novice but sufficiently familiar with the domain? This question is complex and the possible answer seems to depend on the rating context and the objectives pursued. However, in the present context, we would argue that psychologists should be considered as novices because they have no expertise in the artistic domain and their so-called familiarity with the domain is based only on their experience in psychology. If the judges were developmental or educational psychologists, it could have conferred them a certain expertise regarding children's creative skills, but this information was not indicated.

In a third experiment published in the same article, Amabile (1982) asked two kinds of judges to rate creativity of collages made by children from six to eight years old. Alpha coefficients were respectively 0.81 for the artists judges and 0.83 for the non-artists. Furthermore, the correlation between artists and non-artists' ratings was $r = 0.69$. In this research realized in comparable conditions with the precedent one, the intergroup agreement is much higher. Nevertheless, this effect might be caused by the fact that the group of non-artists was composed of undergraduate and graduate students in psychology (i.e., novices) but also by elementary school teachers (i.e., who may be considered as quasi experts). Also, we do not have much information on the characteristics of the artists. Thus one might wonder if the expertise of the artists is certified according to criteria of Kaufman and Baer (2012). If our interpretation is correct, we might conclude that these groups of judges are not sufficiently contrasted to consider one group as novice and another as experts. Once again, the interpretation of such results is complex, even more due to the fact that this study was not designed to systematically compare the ratings from judges with different levels of expertise.

Fortunately, research with more interpretable results exists. For example, Hickey (2001) systematically compared the ratings of different kinds of judges on musical productions. First, she asked children aged from nine to eleven enrolled in music schools to compose short music tracks. Then, these productions were rated by different kinds of judges: three professional composers who had at least 15 years of experience with writing music in a wide variety of genres (composers), college theory professors with at least 10 years' experience in teaching music theory (music theorists), different categories of music teachers (10 "instrumental" music teachers, 4 "mixed-experience" teachers–teachers who taught a combination of instrumental and

choral or instrumental and general music– and 3 "general/choral" music teachers– elementary general music teaching with some choral music), seventh-grade children, and second-grade children. A first important result is that composers showed no consistency in their evaluations ($\alpha = 0.04$). However, according to Kaufman and Baer (2012), composers are the only group of judges that can be considered as truly expert. The interrater reliability of the quasi experts (the music teachers) varies with the type of teaching. The reliability coefficients were relatively acceptable for the general/choral music teachers ($\alpha = 0.81$) and the music theorists ($\alpha = 0.73$). In contrast, the reliability coefficients were less satisfactory for the instrumental music teachers ($\alpha = 0.65$) and the mixed-experience teachers ($\alpha = 0.53$). In this research, younger and older children compared to those who created the music tracks composed the two groups of novices. Not surprisingly, the reliability coefficients were weak for the two groups ($\alpha = 0.50$ for the group composed of 7–8 years old children and $\alpha = 0.61$ for the group composed of 12–13 years old children). Even if these coefficients seem weak, they are equivalent to those obtained by the less consistent groups of teachers. The analysis of the correlations between these different groups of raters showed that three kinds of teachers agreed between themselves but also with the music theorists (inter correlations varied from $r = 0.63$ to $r = 0.88$). Also, the ratings from the two groups of children showed a strong correlation ($r = 0.83$). However, the children's ratings did not correlate well with the teachers' ratings (the strongest observed correlation was only 0.41). This study was interesting because it attempted to study systematically the consistency of ratings from different experts. However, the absence of interrater reliability makes it impossible to aggregate the experts' ratings and to correlate this composite score with that of other raters.

A series of studies made it possible to compare the consistency of ratings from different types of experts. In a first study, Kaufman et al. (2010) asked 205 students to write a small poem and a short story from a given title. The creativity of these two kinds of productions was rated by two types of judges. Poets who had published composed the group of experts and students with no particular skill in the domain composed the group of novices. Regarding poems, with a comparable sample of judges, the experts' interrater reliability was higher than novices' (respectively $\alpha = 0.83$ and $\alpha = 0.57$). The correlation between ratings from the two groups was weak but significant ($r = 0.22$) (Kaufman et al., 2008). For the creativity ratings of the short stories, the interrater reliabilities were comparable to the ones found for poems. However, the correlation between ratings made by poets and students was stronger in this case ($r = 0.71$) (Kaufman et al., 2009). In view of these findings, it seems difficult to conclude that novices are not consistent in their evaluations and that their evaluations do not correlate with the ones made by the experts. These results might also confirm the necessity to select very precisely the expert judges. Indeed, it is also possible to conclude that poets differentiate more with novices when rating poems than when rating short stories because the second type of production was not exactly their field of expertise. Thus, we cannot consider that the large domain of literature can be rated by judges from different literary specialities.

However, for Galati (2015), the scientific debate on expert—novice comparisons has not sufficiently taken into account an important methodological aspect: the variable complexity of the tasks with which raters are confronted. This complexity is defined by the author as "the difficulty to judge something (an idea, a product, a painting, etc.) in function of the particular assessing situation (object's and the judge's characteristics)" (Galati, 2015, p. 25). The complexity of an evaluation is in fact determined by (a) the originality of the product, its appropriateness regarding the context, the complexity of the product itself and its level of diffusion (i.e. the extent of use of the product), and (b) two characteristics of the judges: their experience and their expertise regarding the production domain. The results of Galati's (2015) research on the creativity of paintings showed that novices' ratings had acceptable reliability (0.83) but lower than experts' ratings (0.97). Moreover, Galati asked an expert in the history of art to indicate the complexity for non-experts for rating the creativity level of different paintings. It demonstrates above all that in simple situations, the mean novices' rating did not significantly differ from the experts' mean rating. In contrast, paintings were evaluated as more creative by novices when the evaluation situation was complex.

Based on current scientific knowledge, it seems difficult to conclude on the effect of the different levels of expertise on the subjective evaluation of creativity. Ratings of creative productions by novices can be consistent, and sometimes even superior to domain experts's ratings. Furthermore, we sometimes observe a correlation between experts and novices' ratings but not in every case. These results, in accordance with the Consensual Assessment Technique, are insufficient to render satisfactory novices' ratings of creative productions (Kaufman & Baer, 2012; Kaufman et al., 2009). It seems best to opt for experts' ratings insofar as such experts actually exist and are accessible and willing to participate. However, according to Galati (2015), it is possible to resort to novices' ratings in simple rating situations. On the contrary, it is necessary to select experts in complex situations of evaluation because novices' ratings can be misaligned.

4 Beside the Level of Expertise, What Makes Subjective Evaluations Vary?

The level and type of expertise constitute important sources of evaluation variability, but other personal characteristics that are less studied might also lead to individual differences in creativity evaluation. For example, regarding the creativity level of advertisements, White and Smith (2001) noticed that ratings were significantly correlated with demographical variables (sex and age), reading newspapers and professional experience in the field.

But certain characteristics related to judges' creative potential could also result in individual differences in creativity evaluation. To our knowledge, Hood (1973) was the first author to test this hypothesis. First, he asked participants to indicate as many

unusual uses of a given object as possible. This exercise evaluated the participants' level of originality. Then in a second part, the same participants had to rate the originality of ideas obtained via the same exercise they had to complete previously. The author observed that judges who have less original ideas are more sensitive to variations of productions' originality. Indeed, participants with a moderate or high level of originality discriminated less the variation of originality level and rated the productions more generally as low on originality. These results suggest that judges with a higher level of originality could conceive creativity more narrowly and thus consider that only extremely original productions are creative. Moreover, Caroff and Besançon (2008) in a study on the evaluation of creative advertisements found also the existence of an interaction between judges' levels of originality and their evaluations of productions' creativity, but this interaction showed the opposite effect. Indeed, results indicated that the more judges showed originality in a divergent thinking task, the more they were sensitive to variations of the advertisements' creativity level.

Some authors started to expand their research to further variables. Storme and Lubart (2012) studied how individual differences of intelligence and personality could influence the evaluation of creativity. Their results showed that factor *g* and a personality trait, preference for novelty, were both related to the importance that judges attributed to originality in their evaluation of creativity. In a slightly different perspective, Silvia (2008) studied the effect of personality on people's capacity to discern their own creativity. Participants with a high level of openness in a big five test realized the most creative productions in a divergent thinking task and were also the most exacting when they were asked to select their most creative productions.

To conclude, while seeking to select appropriate judges to rate creativity, it is important to give careful consideration to their type and level of expertise within a specific domain. However, certain variables should also not be neglected, such as the experience or the creative potential related characteristics, because they have an effect on judges' subjective evaluation of creativity.

5 Implicit Conceptions of Creativity as the Base of the Evaluation

Amabile postulates the existence of a common subjective construct of creativity shared by similar judges but she did not seek to understand in detail the nature of this construct (Spiel & von Korff, 1998). The definition of creativity refers to what is in the heads of judges, without specifying their conception or criteria for evaluating it (Katz & Giacommelli, 1982). However, as we highlighted previously that different judges might still agree on creativity ratings, it seems necessary to study implicit conceptions of creativity in order to understand how experts evaluate the creativity of productions.

5.1 Studying Implicit Theories of Creativity

For Runco and Bahleda (1986), implicit theories «are derived from individuals' belief-systems, and are important because they presumably function as a prototype against which (…) behaviors are gauged» (p. 93). Subsequently, Runco and Johnson (2002) developed this definition stating that "implicit theories, from which expectations are formed, are the constellations of thoughts and ideas about a particular construct that are held and applied by individuals" (p. 427). They specified also that these implicit theories are involved, intentionally or not, when we seek to evaluate certain characteristics or behaviors. This idea was developed by Szen-Ziemiańska (2013) who considers that a person will evaluate creativity more or less precisely according to the nature of their implicit conceptions. But implicit theories do not only play a role in the subjective evaluation of creativity. For example, Katz and Giacommelli (1982) supposed that implicit theories might also drive how people will foresee producing something creative. Ultimately, Glăveanu (2014) proposed that implicit theories of creativity presented several common characteristics with social representations.

Few studies have been published on implicit theories of creativity compared to other topics (Ramos & Puccio, 2014; Spiel & von Korff, 1998). A common objective is to extract the experts or novices' implicit conceptions. To do so, authors resort to diverse methodologies. The predominant one consists of an open question. For example, subjects can be asked to write their personal conception of creativity, (Petocz et al., 2009; Spiel & von Korff, 1998; Szen-Ziemiańska, 2013; Tsai & Cox, 2012), to list synonyms of creativity (Ramos & Puccio, 2014; Runco, 1984), behaviors of a creative person (Runco, 1984; Sternberg, 1985), traits that characterize a creative person (Runco, 1984), characteristics of different forms of creativity—artistic, scientific, and daily living (Runco & Bahleda, 1986); to indicate the relation between creativity and a given professional domain (Petocz et al., 2009; Tsai & Cox, 2012), or to imagine a creative product and then describe the characteristics of the person who could have created it (Hass, 2014). However, we can also seek to identify implicit conceptions from standardized material. For example, Katz and Giacommelli (1982) asked researchers to select from the Adjective Check List (Gough & Heilbrun, 1965) the adjectives that described the best the activity of problem-solving. Next, the adjectives were categorized freely by students. Half of the students received the information that the adjectives characterized an activity of problem-solving. For the other half, the adjectives characterized a creative activity which enabled to discriminate specific implicit conceptions of creativity. In a set of experiments on the evaluation of children's creativity, Runco (1989; Runco & Johnson, 2002; Runco et al., 1993) asked a first group of participants to select adjectives from the Adjective Check List (Gough & Heilbrun, 1965) that characterized a creative child. The selected adjectives were then used to build a questionnaire from which a second group of participants rated children's creativity. A third methodological approach consists of analyzing observations and structured interviews of creative persons who were potentially recruited (Elsbach & Kramer, 2003), or to analyze the content of job

offers targeted to select creative persons (Christensen et al., 2014). Finally, there are methodologies used in research on social representations (Glăveanu, 2014).

5.2 *Does the Existing Knowledge on Implicit Conceptions of Creativity Enable Us to Test the Assumptions Behind the CAT?*

The hypotheses formulated by Amabile (1996) state the existence of a rudimentary form of creativity, a basic quality of the product that judges perceive and use to rate the level of creativity. She assumes that this conception should not differ from the scientific conception of creativity that includes two criteria: the levels of originality and adaptation of productions. These assumptions raise two questions. First, do the judges' implicit conceptions of creativity fit the scientific conception? Second, even if the numerous empirical results of Amabile and her research team lead to the conclusion that « the existence of a *unique* subjective construct called « creativity» has been demonstrated[1]» (Amabile, 1983, p. 61) the question is to understand if implicit theories of creativity ascertain a unitary conception of what makes a production creative. Even if it seems very difficult to synthesize our current knowledge, few studies highlight the content of implicit conceptions and how they drive the evaluation of creativity.

Christensen et al. (2014) suggest that research has not suffciently looked at the correspondence between implicit and scientific conceptions of creativity. However, it would be of great interest to retrieve from the implicit conceptions the two criteria of originality and adaptation on which the scientific community bases the study of creativity.

Some research has started to address this issue. Spiel and von Korff (1998) studied implicit conceptions of creativity by asking politicians, scientists, artists and teachers to associate expressions with the word creativity. For these four groups of subjects, the most given expression was "novelty", the second was "idea". Ramos and Puccio (2014) also proposed a free association task with the word "creativity" to two convenient samples. Among the most given answers, we found originality related expressions: New, Unusual, Different and Unique. Szen-Ziemiańska (2013) asked managers and CEOs what they meant by "creativity". From the content of the answers, creativity refers to the aptitude to think creatively, to solve problems by generating new ideas. It is worth noting that there was no expression linked with creativity that refers directly or indirectly to the second scientific criteria of creativity—adaptation.

Even if there is not much research, these scientific results support only partially the hypothesis formulated by Amabile (1996) according to which the conceptual definition of creativity—a production that is both original and adapted—is aligned with

[1] Which we wish *to highlight our point*.

experts' lay conceptions of creativity that is used when evaluating creative productions. If some research has shown that originality is a frequently-cited component of creativity, the criteria of adaptation has never been cited either spontaneously or incidentally.

The CAT has been mostly used to assess the creativity of productions. However, according to Amabile (1996) it can be used under certain conditions to assess individual differences. Thus, it seems relevant to study how a creative person is conceived based on implicit theories of creativity. For example, Katz and Giacommelli (1982) asked their colleagues to select from the Adjective Check List (Gough & Heilbrun, 1965) the adjectives that best described the activity of problem solving. Then, the selected adjectives were freely classified by students so that each category represented an aspect of creativity. The category analysis led the authors to conclude that subjective conceptions of creativity are composed of one dimension of general openness to ideas, situations and actions. Szen-Ziemiańska (2013) obtained an equivalent result. But other studies attempted to conciliate more systematically implicit and scientific conceptions of the creative person. Runco (1984) identified student teachers' stereotypes of a creative person. The expressions resulting from his study were "Flexible", "Non-conforming" and "Challenging" which fit the previous research findings on creative personality. More recently, an original research analyzing job ads conducted by Christensen et al. (2014) highlighted that ads that explicitly sought to recruit creative people feature significantly more terms related to Openness to experience and to a lesser extent Extraversion. Conversely, they feature significantly fewer terms related to Conscientiousness. These results show clearly the correspondence between scientific findings and implicit theories on creative personality.

Moreover, some researchers sought to study more broadly implicit theories of creativity. For example, in a previously cited research Spiel and von Korff (1998) analyzed the participants' answers to determine how the content of implicit theories referred to the "4P" of creativity (Rhodes, 1961). Results indicate that implicit conceptions refer principally to the person or the creative process compared to the product. Moreover, the process is very rarely raised. Furthermore, Szen-Ziemiańska (2013) showed that managers 'conceptions of creativity were globally consistent with scientific theories.

Even if the main objective of the CAT is to provide a subjective but rigorous evaluation of creative productions, these different results establish the applicability of the CAT to evaluate creative people. Indeed implicit conceptions of traits associated with a creative person are consistent with scientific theories. Particularly, openness to experience is in both cases an important determinant of individuals' creative potential. These results strengthen the conviction that appropriate judges' subjective evaluation of creativity offers an alternative solution to evaluation methods based on scientific conceptions.

Finally, can we suppose reasonably that a consistent implicit conception of creativity is more or less shared by equivalent judges? Inversely, do conceptions of creativity vary according to the type of solicited judges (experts, quasi experts or novices) or even among a group of experts in a given domain? If so, the opportunity to aggregate ratings from several judges might be compromised. Implicit theories of

creativity have been collected and studied from different types of potential judges. For example, several studies have been conducted on students (Hass, 2014; Katz & Giacommelli, 1982; Ramos & Puccio, 2014; Runco & Bahleda, 1986), teachers (Runco, 1989; Runco & Johnson, 2002; Runco et al., 1993), and professors in art, business, philosophy, and physics (Sternberg, 1985). To pursue the aim of this chapter, it would be of great interest to find studies comparing implicit conceptions from different judges (experts, quasi experts and novices for example) when confronted with the same experimental design. However, such studies are almost nonexistent. An exception is the notable work of Runco (1989; Runco & Johnson, 2002; Runco et al., 1993) who compared implicit theories on school children's creativity from two groups of judges—parents and teachers. On the 25 adjectives chosen by parents and teachers (from the Adjective Check List, Gough & Heilbrun, 1965), only 7 were common to both groups: Artistic, Curious, Imaginative, Independent, Inventive, Original, and Wide interest (Runco, 1989). Such a low rate of overlap leads us to the conclusion that regarding children's creativity the two groups of "experts" have different implicit conceptions but further research is needed to draw conclusions.

6 What is the Coefficient Alpha Measuring in the Case of Subjective Evaluations of Creativity?

In the classical psychometric approach, we seek to test the validity of a measure. In contrast, in the approach suggested by Amabile (1982), validity is tested by a logical argument stating that: *if* solicited judges are sufficiently familiar with the domain of the creative production (even experts, depending on the criteria we decide to select) and *if* the reliability of their evaluations is high, *then* what experts agree on can only be creativity. However, an implicit assumption underlies this argument: the reliability of evaluations among judges, demonstrated by a high value of the alpha coefficient, is traducing that experts assess collectively *the same characteristic* in different productions. It is indeed tempting to believe that a reliable evaluation of productions by experts is reflecting the level of creativity of these productions. Thus, the evaluation would be valid. But the accuracy of such reasoning is based on the premise that creativity consists of a unique characteristic that is present in every production and that experts recognize it unanimously. Empirically, it means that the reliability coefficient, most of the time estimated by the Cronbach's alpha coefficient, should be interpreted as an indicator of the homogeneity and not only as the internal consistency. Therefore, if our understanding is exact, this reasoning is faulty because it considers that the alpha coefficient allows us to estimate the homogeneity of experts' evaluations regarding creative productions.

A preliminary comment on how authors interpret the value of the reliability index is appropriate before going further in our analysis of Amabile's argumentation. Kaufman et al. (2008) noted that in research using the CAT, the value of inter-rater reliability coefficient ranges from 0.70 to 0.90. For some authors (for example

Hennessey et al., 2011), a reliability index that is at least of 0.70 certifies an acceptable interrater agreement. In fact, the idea that the alpha coefficient should reach 0.70 or 0.80 to conclude satisfactory reliability is widespread among researchers in psychology (Cho & Kim, 2015). Yet, such thresholds have never been supported by empirical testing, psychometric justification nor rational analysis (Churchill & Peter, 1984; Cortina, 1993; Peterson, 1994). In fact, we should avoid concluding mechanically based a simple comparison of the alpha value with some kind of index value (Cho & Kim, 2015). Instead one has to put in more effort to take into account the context and the objective of the evaluation in order to interpret appropriately the alpha (Cortina, 1993).

6.1 When is Alpha a Valid Measure of Reliability?

It is necessary to verify the reliability of ratings. In this way we can ensure that the proportion of error is negligible and that differences of scores reflect the judges' systematic rating of productions (Tinsley & Weiss, 2000). Amabile used different indexes to estimate interrater reliability (Amabile, 1982, 1996), but the alpha coefficient seems to be the most popular lately among researchers using the CAT. Yang and Green (2011) supposed that this preference could be explained because it is an easily interpretable index. Yet, we will see that it is not as easy as it appears.

Generally the alpha coefficient, as any psychometric index, is used to estimate the reliability of a composite score if the hypotheses from which the scores are derived have been respected in empirical conditions. In practice, it is likely that these hypotheses are violated which can skew the empirical estimation of the composite score reliability. Such questions have been extensively studied in the psychometric literature (for example, Cho & Kim, 2015; Cortina, 1993; Green & Yang, 2008; Green et al., 1977; Lucke, 2005; Schmitt, 1996; Sijtsma, 2009; Yang & Green, 2011). Our objective is only to raise issues that we might encounter if we do not respect the validity conditions of this index while we use it to estimate interrater reliability in the case of the subjective evaluation of creativity. We will discuss the appropriateness of this index regarding its utility for the consensual assessment technique.

The well-known assumptions underpinning the alpha coefficient follow the classical theory of composite score reliability that has been calculated from different elementary scores. According to this theory, each elementary score is actually composed of two parts: the true score (for example the real level of creativity of a production) that we seek to estimate, and the measurement error that is supposed to be random. Furthermore, it posits that for each pair of elementary measures, measurement errors must not be correlated. However, since the first work on alpha coefficients (Cronbach, 1951) or equivalent indexes such as the one developed by Guttman (1945), authors conclude that this index provides a lower bound estimation of the real reliability of a composite score. Subsequently, Novick and Lewis (1967, Theorem 3.1) demonstrated that the necessary and sufficient condition for alpha to

really estimate reliability was that every elementary measure would be essentially tau-equivalent which means that for each assessed characteristic, the estimated true scores from two distinct measures are linked by linear functions. Green and Yang (2008) sought to test the importance of the reliability estimation bias from fictive scales that did not respect the presumption of tau-equivalent measures. They verify that the alpha value is always below the reliability value but that this estimation bias stays low (less than 5%) in most of their studied cases. Nevertheless, it can reach 10% of the real reliability value when there are few items and they present very contrasted factor loading values regarding the latent dimension. Most of the time, we observe that the presumption of tau-equivalent measures is not respected in practice (Green & Yang, 2008; Yang & Green, 2011). Thus, the use of alpha underestimates the reliability of measurement scales.

Some research studied the infringement of a second assumption. We pointed out the classical theory on reliability postulates that measurement errors should be random and thus should not correlated. We consider in psychometrics that correlations between errors can occur when subjects do not answer independently to every item composing the test. In other words, when their answers to two items are linked by a second variable that is generally ignored by researchers and that is different from their true score (Lucke, 2005; Raykov, 2001). Numerous reasons have been evoked to explain correlations between errors (Cho & Kim, 2015; Green & Yang, 2008; Lucke, 2005; Yang & Green, 2011). This is certainly why little attention has been paid to consequences of the infringement of this presumption on the reliability estimation (Green et al., 1977; Lucke, 2005), even if this bias is well-known since the article of Guttman (1953). But whatever the reason is to explain correlations between errors, these correlations should skew the calculation of the alpha because the covariance between errors in taken into account in the calculation of the mean covariance between items, which appears in the numerator of alpha.[2] This bias has been highlighted in different studies. Analyses show that alpha overestimates the reliability when covariance between errors is positive (Raykov, 1998, 2001) and underestimates it when the covariance is negative (Raykov, 2001). The effect of correlated errors on alpha had been subject of simulation studies conducted by Lucke (2005) then Cho and Kim (2015). Respectively they found biases in the reliability of congeneric measures and on the alpha value. Indeed the more the measurement errors are correlated, the more the measurement reliability decreases while at the same time alpha tends to overestimate reliability.

[2] The definition formula proposed by Cronbach (1951; Eq. 2) is well-known: $\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right)$, where n is the number of items, V_i is the variance of an item i and V_t is the variance of the composite score. From this equation, the author derived two other formulas: $\alpha = \frac{n}{n-1}\left(\frac{\sum_i \sum_j C_{ij}}{V_t}\right)$ where C_{ij} is the covariance between each different pairs of items ($i \neq j$) and $\alpha = \frac{n^2 \bar{C}_{ij}}{V_t} where \bar{C}_{ij}$ is the covariance between every items (Cronbach, 1951; Eqs. 24 and 16 respectively).

6.2 *How Should We Interpret Alpha When Estimating Interrater Reliability?*

Infringing on the two assumptions presented earlier is not the only risk that affects the interpretation of the coefficient and the conclusions to be drawn. The value of alpha depends also on empirical conditions from which it is estimated. It has been argued that the value of alpha for a composite score varies according to different parameters: the number of items, their mean correlation or even the number of dimensions that are truly measured by the items. We will address successively the effect of these different parameters and their impact when the alpha coefficient is used to estimate interrater reliability.

The relation between alpha and the number of measures composing the test is familiar to psychometricians and known among researchers using the subjective evaluation of creativity. Amabile (1982; note 2, p. 1003) stated herself: « to the extent that the judging is a difficult task and the mean inter-judge correlation is low, the number of judges should be increased. However, if the mean correlation is high, good reliabilities can be obtained with fewer judges». This idea had been further promoted by Kaufman et al. (2008), for whom the more the number of judges asked to assess creativity is high, the more the interrater reliability has chances to be high. According to them, an optimal number of judges should lie between five and ten in most evaluation situations. Resorting to less than five judges means taking the risk to obtain insufficient reliability. On the other hand, seeking to obtain more than ten judges might often be unnecessary and costly. Aside from the fact that this recommendation is certainly too general, none of the thresholds given were justified by the authors. Thus we cannot draw appropriate conclusions as to the interpretation of alpha regarding the consensual assessment technique.

We can rely on formulas derived from Cronbach (1951) to demonstrate that the value of the index depends on the number of items but also the mean correlation between items.[3] Green et al. (1977) were the first to analyze the effect of the number of items on alpha. Results from their Monte Carlo study attest the increase of alpha when the number of items composing the scale increases. Subsequently, Cortina (1993) demonstrate that on the one hand increasing the number of items enhanced considerably the value of alpha, particularly when the mean correlation between items was weak (i.e. 0.30), and on the other hand that if the scale contains sufficient items (i.e. more than twenty) the value of alpha exceeded 0.70 even when the mean correlation between items was weak (i.e. 0.30). Therefore, if internal consistency means that items composing a test are interrelated (Cortina, 1993; Green et al., 1977; Schmitt, 1996; Sijtsma, 2009), we cannot conclude only from the value of alpha because this index depends on mean correlation between items (i.e. their interrelation) and on the length of the test. A correct interpretation of alpha implies that the number of items and the mean correlation between items be taken into account simultaneously. When this index is used to evaluate the reliability of subjective ratings of

[3] See previous footnote.

creativity, we should cautiously take Kaufman et al. (2008) recommendations to have five to ten judges, as well as Hennessey et al. (2011) who consider that reliability is acceptable from the threshold of 0.70. How should interrater reliability be interpreted in such conditions? A part of the answer can be found thanks to the following example. From a derived formula of alpha (Peterson, 1994), we can estimate that a coefficient of 0.70 calculated from evaluations of ten judges should correspond to a mean correlation of 0.19 which obviously translates a weak consistency between individual evaluations (in the exact same condition, an alpha value of 0.80 would correspond to a mean correlation of 0.29).

We mentioned that following the CAT, the interrater reliability is conceived tacitly as the index of homogeneity among subjective evaluations, meaning that every judge is evaluating the same characteristic. Yet, if alpha cannot be directly interpreted as an index of internal consistency, neither can it be interpreted as an index of homogeneity. In psychometrics, if a common factor for every measure exists then it leads necessarily to an index of high internal consistency. But the opposite is not true. It is possible to obtain a high alpha even when more than a common factor can explain the score variance, including orthogonal factors (Cortina, 1993; Green & Yang, 2008; Green et al., 1977; Sijtsma, 2009). In fact, an analysis demonstrates that the value of alpha varies depending more on the increase of the explained variance for each score than the structure of the measure itself (the number of more or less important of factors and their relations) (Cho & Kim, 2015; Green & Yang, 2008). Acknowledging this should lead us to question the dimensionality of subjective evaluations of creativity. Using alpha to estimate interrater reliability would lead to skewed results if we have any reason to think that judges will base their evaluation of creativity on more than one dimension. These dimensions can incorporate, as stated by Amabile (1996) creativity and one or more other dimensions that are common to every production. But these dimensions can also not refer to creativity or represent different facets of creative production (for example originality and adaptation).

To conclude, the direct interpretation of the value of alpha should be avoided. We should analyze and interpret this index by taking into account different parameters of the CAT: number of solicited judges, the mean correlation between their evaluations and the dimensionality of the obtained measures of creativity.

7 Conclusion and Perspectives

This chapter sought to address three general questions related to the use of the CAT. What type of judges should be selected to rate creative productions? What implicit conceptions on creativity are guiding the judges' evaluations? And, does the alpha coefficient give an adequate estimate of the reliability of the evaluations? In other words, we aimed to discuss how valid can be considered the results from the consensual assessment technique. The analysis of the numerous results in the literature leads us to several conclusions.

- The original hypothesis on which the CAT is based is that a subjective construct of creativity exists and is common to different judges as long as they have an equivalent level of expertise in a given domain (Amabile, 1996). Research results led us to think judges assess creativity based on implicit conceptions that are not as reliable as we could wish. Moreover, it has not been demonstrated yet that these implicit conceptions fit with scientific theories and conceptions.
- Regarding the selection of judges, experts in the domain seem obviously in the best position to assess the creativity of productions and should be favored every time it is possible. However if we want to stick with strict criteria, for example Kaufman and Baer's criteria (2012), it is often impossible to access judges who have at least ten years of specific experience within the domain and have received an honor for their exceptional realizations within the same domain. In fact, expert judges are not necessarily required in every case. It depends on the productions that we wish to assess. Results show that novices were able to assess efficiently creative productions in certain conditions. Thus we can conclude that the selection of judges should be guided by pragmatic considerations.
- The use of the evaluation technique itself should strictly follow the methodological principles that have been outlined by Amabile (1982, 1996; Hennessey et al., 2011). However, we observe that two important principles are forgotten in numerous studies. First, we deplore that judges are most of the time asked to evaluate each production directly on a rating scale (such as "not creative at all"—"very creative"). Instead they should be asked to compare and rate each production related to the others. Second, we need to ask judges to rate other aspects of the productions (such as technical and aesthetic qualities, or other aspects). These recommendations are rarely put into practice when using the CAT. However, they should be applied systematically. By doing so, we will be in position to establish without ambiguity that judges are indeed rating creativity independently from other related and assessed aspects that may be taken into account or confounded when rating creativity.
- The estimations of rating reliability cannot be established by interpreting directly the alpha coefficient. Even if this statistical technique is actually privileged by a majority of researchers, we should be cautious and take into account different parameters that influence the value of the alpha. Indeed, we recall that depending on the number of solicited judges, the mean correlation between ratings and the dimensionality of creativity measures can lead to an overestimation or an underestimation of the interrater reliability. This bias varies in its proportion and might not be easily identifiable. Given this, we should reconsider using the inter class correlation coefficient previously adopted by Amabile (1982, 1996).
- Finally, a more or less high interrater agreement cannot by itself lead to conclude about the validity of the creativity rating. Theoretically, we cannot definitively exclude that judges (even experts) may rate productions coherently but on a different construct than creativity. Also, their creativity ratings might be contaminated by other characteristics of the productions. This argument is even more crucial given that, as stated before, most researchers do not control that creativity ratings differ from related aspects (esthetical or technical qualities) when using the

CAT. The empirical validity of the creativity evaluation needs to be approached from a different perspective.

If one statement is to be remembered, it is the difficulty to rely blindly on experts ratings, even when reliable, to consider that this technique leads to a valid evaluation of creativity.

The current approach to test empirical validity based only on a statistical agreement between judges' ratings is problematic because we do not seek to understand what are the judges' implicit conceptions of creativity, if they are coherent between judges, or how such implicit conceptions guide judges' evaluations. According to Runco and Johnson (2002), one limit of research on implicit theories is the tendency to remain descriptive rather than seeking to explain behaviors. Thus, they suggest « to study the conceptions of creativity in conjunction with the observed behaviors of their application» (Runco & Johnson, 2002, p. 437). Symmetrically, we consider that the subjective assessment of creativity would gain in comprehension and the evaluations in validity if the judgments were confronted to judges' implicit conceptions. Amabile (1996) conducted a first study of this kind. She aimed to verify that judges perceived a creative production as both new and adapted. To do so, judges had first to rate creative productions then to answer an open question in order to describe their subjective impressions. But the results were deceiving: answers were unclear, difficult to analyze and presented a high degree of variability.

However, we are convinced that it is possible to enhance the validity of this technique by modifying it in at least two ways. Runco (1984) was the first to offer an alternative to evaluate creativity. Runco shares with Amabile (1982, 1996) the desire to develop a socially valid instrument to evaluate creativity, but his approach consists, first, of collecting implicit conceptions from adequate people regarding the production under evaluation, and then based on these implicit theories, to *construct* an instrument of evaluation (for example, Runco, 1984; Runco & Bahleda, 1986). This approach should guarantee the instrument better ecological validity than if it was constructed based only on scientific theories (Runco, 1984; Sternberg, 1985). Concretely, a first strategy consisted of asking teachers to list expressions that were synonyms of creativity, observed behaviors that are specific to creative children and personality traits that would be common to all of them (Runco, 1984). The most frequently cited items were retained for the questionnaire assessing children's creativity. It is important to identify the most appropriate people to collect their implicit theories. If not, we may obtain different items from different types of experts (Runco & Bahleda, 1986). A second strategy consists of updating an approach previously used by Domino (1970) and Gough (1979). They asked parents and teachers to select among the 300 items of the Adjective Check List (ACL; Gough & Heilbrun, 1980) the traits that were, in their opinion, related to children's creativity (Runco, 1989; Runco et al., 1993). Once again, the questionnaire to assess creativity was constructed using the most frequently cited adjectives. Existing results tend to suggest good validity for this technique (Runco, 1989; Runco & Bahleda, 1986; Runco & Johnson, 2002; Runco et al., 1993). Inspired by these examples, it is certainly possible to develop other subjective evaluation techniques of creativity that are socially valid.

A second approach would consist of transgressing a powerful interdiction resulting from a certain methodological orthodoxy. Stating that judges should rate creativity according to their own conception of such dimensions translates the interdiction to provide us with any kind of definition or criteria to use in order to rate the productions. This injunction results from the statement made by Amabile (1982, 1996) that it is not possible to specify to which objective characteristics of the productions creativity corresponds. For this reason she decided to dissociate conceptual and operational definitions of creativity. But while doing so, she also postulated that the consensual assessment of creativity made by judges should certainly rely on the two dimensions stipulated in the standard definition of creativity: novelty and the appropriateness of productions. In opposition to her recommendations, we think that we could ask judges to rate these two criteria according to their own conceptions, without giving them indications. By doing so, we would request their expertise to rate the two dimensions that constitute creativity rather than asking them a subjective and implicit evaluation of creativity. Thus we would be able to align conceptual and operational definitions of creativity.

Finally, regardless of the different issues highlighted in this chapter, another question remains unanswered. We still do not know how judges evaluate the creativity of a production. Amabile (1996) herself pointed out the necessity for complementary research in order to understand which characteristics of the judgment task and the judges themselves might influence interrater agreement. Hennessey (1994) called for studying the differences between judges for themselves, and not only to seek to enhance ratings' reliability. In the same vein, Runco and Charles (1993) called for research exploring how judges proceed to rate the creativity of productions. Indeed, we need to verify that judges effectively take into account the relevant dimensions of productions. Then, we will be able to seek to understand how they integrate the adaptation dimension to the originality dimension when giving a global judgment of creativity.

Bibliography

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013. https://doi.org/10.1037//0022-3514.43.5.997

Amabile, T. M. (1983). A consensual technique for creativity assessment. In *The social psychology of creativity* (pp. 37–63). Springer.

Amabile, T. M. (1996). *Creativity in context*. Westview Press.

Baer, J. (1996). Does artistic creativity decline during elementary school? *Psychological Reports, 78*(3), 927–930. https://doi.org/10.2466/pr0.1996.78.3.927

Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal, 10*(1), 25–31. https://doi.org/10.1207/s15326934crj1001_3

Baer, J. (2003). The impact of the core knowledge curriculum on creativity. *Creativity Research Journal, 15*(2–3), 297–300. https://doi.org/10.1080/10400419.2003.9651422

Baer, J., & Kaufman, J. C. (2019). Assessing creativity with the consensual assessment technique. In I. Lebuda & V. P. Glăveanu (Eds.), *The Palgrave handbook of social creativity research.* Springer International Publishing. https://doi.org/10.1007/978-3-319-95498-1

Baer, J., Kaufman, J., & Gentile, C. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*(1), 113–117. https://doi.org/10.1207/s15326934crj1601_11

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). Information Science Reference.

Besemer, S. P. (1998). Creative product analysis matrix: Testing the model structure and a comparison among products—Three novel chairs. *Creativity Research Journal, 11*(4), 333–346. https://doi.org/10.1207/s15326934crj1104_7

Besemer, S. P. (2000). To buy or not to buy: Predicting the willingness to buy from creative product variables. *Korean Journal of Thinking and Problem Solving, 10*(2), 5–18.

Besemer, S. P., & O'Quin, K. (1986). Analyzing creative products: Refinement and test of a judging instrument. *The Journal of Creative Behavior, 20*(2), 115–126. https://doi.org/10.1002/j.2162-6057.1986.tb00426.x

Besemer, S. P., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal, 12*(4), 287–296.

Besemer, S. P., & Treffinger, D. J. (1981). Analysis of creative products: Review and synthesis. *The Journal of Creative Behavior, 15*(3), 158–178. https://doi.org/10.1002/j.2162-6057.1981.tb00287.x

Caroff, X., & Besançon, M. (2008). Variability of creativity judgments. *Learning and Individual Differences, 18*(4), 367–371. https://doi.org/10.1016/j.lindif.2008.04.001

Carson, S. (2006, April 19). *Creativity and mental illness.* Invitational panel discussion hosted by Yale's mind matters Consortium, New Haven, CT.

Chen, C., Kasof, J., Himsel, A., Dmitrieva, J., Dong, Q., & Xue, G. (2005). Effects of explicit instruction to "be creative" across domains and cultures. *The Journal of Creative Behavior, 39*(2), 89–110. https://doi.org/10.1002/j.2162-6057.2005.tb01252.x

Chen, C., Kasof, J., Himsel, A. J., Greenberger, E., Dong, Q., & Xue, G. (2002). Creativity in drawings of geometric shapes: A cross-cultural examination with the consensual assessment technique. *Journal of Cross-Cultural Psychology, 33*(2), 171–187. https://doi.org/10.1177/0022022102033002004

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Christensen, B. T., Drewsen, L. K., & Maaløe, J. (2014). Implicit theories of the personality of the ideal creative employee. *Psychology of Aesthetics, Creativity, and the Arts, 8*(2), 189–197. https://doi.org/10.1037/a0036197

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*(4), 360–375. https://doi.org/10.2307/3151463

Čorko, I., & Vranić, A. (2004). Effects of setting creative goals of different specificity on judged creativity of the product. *Review of Psychology, 11*(1–2), 67–73.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

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*(2), 159–166. https://doi.org/10.1037/aca0000220

Domino, G. (1970). Identification of potentially creative persons from the adjective check list. *Journal of Consulting and Clinical Psychology, 35*(1), 48–51. https://doi.org/10.1037/h0029624

Elsbach, K. D., & Kramer, R. M. (2003). Assessing creativity in Hollywood pitch meetings: Evidence for a dual-process model of creativity judgments. *Academy of Management Journal, 46*(3), 283–301. https://doi.org/10.2307/30040623

Galati, F. (2015). Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity. *Creativity Research Journal, 27*(1), 24–30. https://doi.org/10.1080/10400419.2015.992667

Gerrard, L. E., Poteat, G. M., & Ironsmith, M. (1996). Promoting children's creativity: Effects of competition, self-esteem, and immunization. *Creativity Research Journal, 9*(4), 339–346. https://doi.org/10.1207/s15326934crj0904_5

Glăveanu, V. P. (2014). Revisiting the "art bias" in lay conceptions of creativity. *Creativity Research Journal, 26*(1), 11–20. https://doi.org/10.1080/10400419.2014.873656

Gough, H. G. (1979). A creative personality scale for the adjective check list. *Journal of Personality and Social Psychology, 37*(8), 1398–1405. https://doi.org/10.1037/0022-3514.37.8.1398

Gough, H. G., & Heilbrun, A. B. jr. (1965). *The adjective check list manual*. Consulting Psychologists Press.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient Alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*(4), 827–838. https://doi.org/10.1177/001316447703700403

Green, S. B., & Yang, Y. (2008). Commentary on coefficient Alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282. https://doi.org/10.1007/BF02288892

Guttman, L. (1953). Reliability formulas that do not assume experimental independence. *Psychometrika, 18*(3), 225–239. https://doi.org/10.1007/BF02289060

Hass, R. W. (2014). Domain-specific exemplars affect implicit theories of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 8*(1), 44–52. https://doi.org/10.1037/a0035368

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal, 7*(2), 193–208. https://doi.org/10.1080/10400419409534524

Hennessey, B. A., Amabile, T. M., & Mueller, J. S. (1999). Consensual assessment. *Encyclopedia of Creativity, 1*, 347–359.

Hennessey, B. A., Amabile, T. M., & Mueller, J. S. (2011). Consensual assessment. In M. A. Runco & S. Pritzker (Eds.), *Encyclopedia of creativity* (2nd ed., Vol. 1, pp. 253–260). Academic Press.

Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education, 49*(3), 234–244. https://doi.org/10.2307/3345709

Hood, R. W. (1973). Rater originality and the interpersonal assessment of levels of originality. *Sociometry, 36*(1), 80–88. https://doi.org/10.2307/2786283

Joussemet, M., & Koestner, R. (1999). Effect of expected rewards on children's creativity. *Creativity Research Journal, 12*(4), 231–239. https://doi.org/10.1207/s15326934crj1204_1

Kasof, J., Chen, C., Himsel, A., & Greenberger, E. (2007). Values and creativity. *Creativity Research Journal, 19*(2–3), 105–122. https://doi.org/10.1080/10400410701397164

Katz, A. N., & Giacommelli, L. (1982). The subjective nature of creativity judgments. *Bulletin of the Psychonomic Society, 20*(1), 17–20. https://doi.org/10.3758/BF03334789

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal, 24*(1), 83–91. https://doi.org/10.1080/10400419.2012.649237

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). The consensual assessment technique. In *Essentials of creativity assessment* (pp. 52–83). Wiley.

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise domains and the consensual assessment technique. *The Journal of Creative Behavior, 43*(4), 223–233. https://doi.org/10.1002/j.2162-6057.2009.tb01316.x

Kaufman, J. C., Baer, J., Agars, M. D., & Loomis, D. (2010). Creativity stereotypes and the consensual assessment technique. *Creativity Research Journal, 22*(2), 200–205

Lucke, J. F. (2005). "Rassling the Hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement, 29*(2), 106–125. https://doi.org/10.1177/0146621604272739

Niu, W., & Sternberg, R. (2001). Cultural influences on artistic creativity and its evaluation. *International Journal of Psychology, 36*(4), 225–241. https://doi.org/10.1080/00207590143000003

Novick, M., & Lewis, C. (1967). Coefficient Alpha and the reliability of composite measurements. *Psychometrika, 4866*(1), 1–13. https://doi.org/10.1007/BF02289400

O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal, 2*(4), 267–278. https://doi.org/10.1080/10400418909534323

O'Quin, K., & Besemer, S. P. (2006). Using the creative product semantic scale as a metric for results-oriented business. *Creativity and Innovation Management, 15*(1), 34–44. https://doi.org/10.1111/j.1467-8691.2006.00367.x

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient Alpha. *Journal of Consumer Research, 21*(2), 381. https://doi.org/10.1086/209405

Petocz, P., Reid, A., & Taylor, P. (2009). Thinking outside the square: Business students' conceptions of creativity. *Creativity Research Journal, 21*(4), 409–416. https://doi.org/10.1080/10400410903297998

Ramos, S. J., & Puccio, G. J. (2014). Cross-cultural studies of implicit theories of creativity: A comparative analysis between the United States and the main ethnic groups in Singapore. *Creativity Research Journal, 26*(2), 223–228. https://doi.org/10.1080/10400419.2014.901094

Raykov, T. (1998). Coefficient Alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375–385. https://doi.org/10.1177/014662169802200407

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69–76. https://doi.org/10.1177/01466216010251005

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan, 42*(7), 305–310.

Runco, M. A. (1984). Teachers' judgments of creativity and social validation of divergent thinking tests. *Perceptual and Motor Skills, 59*(3), 711–717. https://doi.org/10.2466/pms.1984.59.3.711

Runco, M. A. (1989). Parents' and teachers' ratings of the creativity of children. *Journal of Social Behavior & Personality, 4*(1), 73–83. https://doi.org/10.1207/S15326934CRJ1434_12

Runco, M. A., & Bahleda, M. D. (1986). Implicit theories of artistic and everyday creativity. *Journal of Creative Behavior, 20*(2), 93–98. https://doi.org/10.1002/j.2162-6057.1986.tb00423.x

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences, 15*(5), 537–546. https://doi.org/10.1016/0191-8869(93)90337-3

Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *Journal of Creative Behavior, 39*(2), 137–148. https://doi.org/10.1002/j.2162-6057.2005.tb01255.x

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Runco, M. A., & Johnson, D. J. (2002). Parents' and teachers' implicit theories of children's creativity: A cross-cultural perspective. *Creativity Research Journal, 14*(3–4), 427–438. https://doi.org/10.1207/S15326934CRJ1434_12

Runco, M. A., Johnson, D. J., & Bear, P. K. (1993). Parents' and teachers' implicit theories of children's creativity. *Child Study Journal, 23*(2), 91–113. https://doi.org/10.1207/S15326934CRJ1434

Schmitt, N. (1996). Uses and abuses of coefficient Alpha. *Psychological Assessment, 8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 139–146. https://doi.org/10.1037/1931-3896.2.3.139

Spiel, C., & von Korff, C. (1998). Implicit theories of creativity: The conceptions of politicians, scientists, artists and school teachers. *High Ability Studies, 9*(1), 43–58. https://doi.org/10.1080/1359813980090104

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49*(3), 607–627. https://doi.org/10.1037//0022-3514.49.3.607

Storme, M., & Lubart, T. (2012). Conceptions of creativity and relations with judges' intelligence and personality. *The Journal of Creative Behavior, 46*(2), 138–149. https://doi.org/10.1002/jocb.10

Szen-Ziemiańska, J. (2013). Psychometric and self-rated creativity of Polish managers: Are implicit theories of creativity relevant to self-assessment? *The International Journal of Creativity & Problem Solving, 23*(1), 59–69.

Tinsley, H. E., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). Academic Press.

Tsai, K., & Cox, S. (2012). Business students' beliefs about creativity. *Journal of Business, 4*(2), 1–10.

White, A., & Smith, B. L. (2001). Assessing advertising creativity using the creative product semantic scale. *Journal of Advertising Research, 41*(6), 27–34.

Yang, Y., & Green, S. B. (2011). Coefficient Alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377–392. https://doi.org/10.1177/0734282911406668