A Scattered CAT: A Critical Evaluation of the Consensual Assessment Technique for Creativity Research

Genevieve M. Cseh[1] & Karl K. Jeffries[2]

Authors' Note

[1]Department of Psychology, Buckinghamshire New University, High Wycombe, United Kingdom

[2]School of Art, Design & Fashion, University of Central Lancashire, Preston, United Kingdom

Correspondence regarding this article should be addressed to Dr. Genevieve Cseh, Department of Psychology, School of Human & Social Sciences, Buckinghamshire New University, Queen Alexandra Road, High Wycombe HP11 2JZ, United Kingdom. Tel: +44 (0)1494 522141 ext 4314, Email: genevieve.cseh@bucks.ac.uk

Abstract

Amabile's *Consensual Assessment Technique* (CAT) – taking the consensus opinions of domain experts – is considered a 'gold standard' of creativity assessment for research purposes. While several studies have identified how specific procedural choices impact on the CAT's reliability as a measure, researchers' depth of knowledge about procedures and their effects still remains incomplete. This paper explores gaps in the research by reviewing CAT and creativity literature, and aims to explore to what extent the creativity research community needs to revisit and reflect on the CAT and solidify protocols for its implementation. The conclusion highlights the need for new debate and a program of research to clarify, evidence, and harmonize CAT methodology, while simultaneously preserving the CAT's flexibility. This would enable the development and sophistication of the CAT, including possible new assistive technologies, to further strengthen its use within the science of creativity.

Keywords: *creativity assessment, consensual assessment technique, research methodology, rating procedures, research protocols*

Since Guilford's renowned 1950 address to the American Psychological Association, creativity research has grown considerably and so have methods to assess creativity (Batey & Furnham, 2006). Creativity has been explored from many perspectives, but frequently focuses on assessing an end product, with definitions of what makes a product creative usually centering on originality and appropriateness (also similarly termed usefulness, value, or effectiveness, amongst others). Yet, such classifications are contested; settling on a definition that suits every instance or form of creativity in every domain is still the subject of much debate (Runco & Jaeger, 2012).

A resolution to this definition debate is to accept the inherent subjectivity of judging creativity and take a theoretically neutral stance on criteria. This approach was operationalized as an assessment method for research over 35 years ago by Teresa Amabile as the *Consensual Assessment Technique* (Amabile, 1982, 1996). It has been espoused as a, if not *the* "gold standard" (Carson, 2006, as cited in Baer & McKool, 2009, p. 2) in creativity assessment, and considered a versatile, reliable measure of creativity negating the need for explicit definitions. After over three decades of its use in creativity research, it is time to reflect on the CAT's contributions to creativity assessment, its strengths and weaknesses, and its future.

What is the CAT?

Amabile (1982, 1996) argued that the concept of what is creative is largely shared amongst a domain's experts as tacit knowledge, and that creativity should therefore be assessed by consensus between domain experts. If sufficient agreement was reached, this would define the level of the product's creativity, relative to the other products within a sample, within a particular context of time and place. This not only eschews and to some degree resolves the longstanding creativity definition debate, but simultaneously quantifies and operationalizes its assessment for the purposes of scientific research.

The CAT can be broadly stratified into three distinct stages of procedure. First, a group of creative outputs from the same or a similar task are gathered. Second, these works are given to a number of suitable judges who each rate the individual outputs on their relative creativity compared to the group of outputs as a whole. This involves a number of protocols such as: briefing the judges on the task; showing them the work; deciding on the rating scale; recording and collecting ratings. Third, the ratings of judges are analyzed to compute the level of consensus, and with a satisfactory level of agreement the outputs can be arranged on a spectrum from lower to higher-scoring creativity. On this basis, relationships can be explored between creativity and other relevant study variables.

CAT use has been growing exponentially. It has been widely used across many disciplines and settings, all educational levels (from kindergarten to higher education), and numerous professions, both those traditionally associated with creativity, and in less obvious domains, such as the military (McClary, 2009).

The Problem: Consistency and Transparency of Methodology

One of the biggest strengths of the CAT is its (seemingly) simple method, and its adaptability to a wide variety of domains. However, as with any scientific tool, standards of compliance, consistency, and transparency are paramount for integrity, replication, and comparison across findings. For Amabile and her colleagues the reliability and validity of the measure was conditional on following certain guidelines (Amabile, 1982, 1996; Hennessey, 1994; Hennessey, Amabile, & Mueller, 2011), for example, that judges should:

- have domain experience;
- rate creativity independently and subjectively, i.e., without new training, discussion, or specific guidance;
- rate creativity relative to a specific sample and context;
- each see the items they are to rate in different random orders (to avoid order effects);

- when assessing a task for the first time, rate factors other than overall creativity (e.g., technical execution and aesthetic appeal), and to use factor analysis to ensure discriminant validity of the creativity measure.

While these guidelines are to a degree specific, they also leave room for interpretation and expansion. Baer, Kaufman, and Gentile (2004) and Christiaans and Venselaar (2005), were able to show that CAT inter-rater reliability could remain high even when creative outputs were not generated under experimental conditions, demonstrating that although the original CAT guidelines offer a foundation on which to build, there is still much to be learned about the methodology's elasticity.

Given the CAT's adaptability to a variety of circumstances inherently necessitates flexibility, some aspects of CAT study design do require more scrutiny, in terms of how they might affect the integrity of results. Likewise, the level of detail in reporting CAT methodology varies, impacting opportunities for replication and cross-study comparisons. Although many CAT studies show the ratings to be reliable, others are less convincing, suggesting that inconsistency of method may also be leading to inconsistency in ratings. The broad conclusion is that these omissions in reporting procedural detail, uncertainties, and inconsistencies are scientifically problematic.

Overview and Aims

Each of the stages of a CAT study procedure present researchers with decisions for interpretation. How many creative outputs will be acquired, and under what conditions? How many judges will be recruited? Who are suitable judges? What instructions will be given? What rating scale will be used? What happens if the level of inter-rater reliability amongst judges is low? The answers to these questions are important, and to a large extent determine the credibility of a study's findings and the CAT's credibility as a measure.

This article does not seek to define what the answers to those questions are; instead it attempts to make the case that there is a need for creativity researchers to further the debate on what constitutes best practice regarding CAT protocol. To this end, this article reviews some of the varied choices researchers in CAT studies have made. The likelihood is various choices of protocol are interrelated, potentially with cumulative influences. In this regard, the task, the creative outputs, the judges, and the rating protocols form a complex system. To explore the complexities of this system it is useful first to discuss individual points of decision-making in the design of a CAT study.

Specifically, this review aims to illustrate some of the variability within: criteria by which judges are selected (expertise); the number of judges recruited; task selection and its intersection with judges' experience; stimuli presentation; rating procedures, including judge instructions, rating scale, and factors measured; statistical analysis techniques and inter-rater reliability. We examine each of these issues in turn with examples of studies that demonstrate the wide methodological range to date. Cited studies are not meant to single out individual researchers or to suggest right or wrongdoing; they are also not exhaustive. They serve as illustrations of the variety of methodological considerations facing all researchers, which require a more in-depth understanding within the creativity research community as a whole.

Suitability of Judges

"…A product or response is creative to the extent that appropriate observers independently agree it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated" (Amabile, 1982, p.1001). Amabile (1982, 1996) has suggested that learned standards specific to each domain are acquired over time through education and personal experience and thereby internalized as shared knowledge of the domain's history and culture, against which new creative offerings

can be judged in context. CAT judges are therefore required to be 'knowledgeable' or 'experts' in the domain to which the task is associated, and "the validity of the CAT is grounded in the fact that experts in a domain are the final arbiters of what is creative (or otherwise valued) in a domain" (Kaufman, Baer, et al., 2008, p.175). Furthermore, "judges should be closely familiar with works in the domain, at least at the level of those being produced by the subjects" (Amabile, 1996, p. 73), while Kaufman, Plucker, et al. (2008) suggest that: "Judges should have a level of expertise that is clearly higher than the presumed level of expertise of the subjects creating the artifacts" (p.74). For example, collage tasks have been rated by artists (Amabile, 1982), poetry by poets (Kaufman, Baer, et al., 2008), music composition by music teachers (Byrne, MacDonald, & Carlton, 2003), and graphic designs by professional graphic designers (Jeffries, Zamenopoulos, and Green, 2017).

Perhaps unsurprisingly, domain-matched judges have been shown to achieve higher inter-rater reliability than unmatched judges. In Amabile's 1982 study in which psychologists, art teachers, and artists rated collages created by children, for example, although all three judge groups showed inter-rater agreement at above .70 (established as the minimum required reliability level to be considered acceptable agreement: Amabile, 1996; Kline, 2000), psychologist judges had the lowest agreement (.73), while art teachers had the highest agreement (.88). On the other hand, Baer, Kaufman, and Riggs (2009) found high reliability and agreement amongst matched and non-matched judges, i.e., psychologists, teachers and writers.

Due to the relative ease with which the general public or university students can be recruited compared to domain experts – who tend to be relatively scarce, busy, and expensive – there is a clear advantage if CAT researchers do not necessarily need highly expert judges. Previous CAT studies have explored if it is appropriate to replace experts with non-experts, and found in comparison to expert raters, novices do not likely yield a sufficient level of

consensus or correlation with expert ratings to act as appropriate substitutes. Kaufman, Baer, et al. (2008) found that novices (106 college students) seemed to show slightly higher agreement than 10 expert poets (.94 compared to the poets' .83) when evaluating the overall creativity of over 200 poems. However, when adjusting for the high novice judge sample size (reducing sample size to 10 for each group), agreement between the novices dropped to .53. A consideration to note is the sensitivity of traditional methods of assessing inter-rater agreement such as Cronbach's alpha to the number of items rated/judges (for more detailed discussion of issues surrounding the statistical analysis of CAT ratings, see our later section on 'Reliability, Agreement, and Statistical Test Choice', and Myszkowski and Storme's paper, also in this issue, on Judge Response Theory).

In contrast, other researchers have not found experts to always produce significantly more reliable agreement than novices. Freeman, Son, and McRoberts (2015) compared the ratings of three expert and three novice judges in rating fashion illustrations using the CAT and found no difference between the expertise groups in terms of agreement. Amabile (1982) noted that in some of her studies expertise did not seem to increase inter-rater reliability as much as expected, noting "no clear superiority of artists over nonartists in interjudge reliability" (p. 1006), while nonartist and artist judges, when rating collages, showed reasonably good inter-rater agreement. Amabile acknowledges that this may be due to the nature of the task being judged – collages are perhaps simple enough that most people have some minimal familiarity with them, whereas more specialist domains (e.g., medicine) likely require much more specialized experience. Likewise, this is a point raised by Kaufman, Baer, and Cole (2009), who found that the agreement between novices and experts was higher for short stories than poetry, indicating a domain or task interaction.

Some argue that a compromise may be struck by using 'quasi-expert' judges – those with intermediate knowledge of a domain, but more easily available – often graduate

students. Kaufman, Gentile, and Baer (2005) found that gifted novice writers were equally matched to more expert writers in their agreed assessment of others' creative poems and short stories, and that there was a good correlation between the ratings of these gifted novices and experts' ratings. Similarly, Kaufman, Baer, Cropley, Reiter-Palmon, and Sinnett (2013), in two studies of judges with varying levels of expertise, rating outputs from a writing and an engineering product design task, also found that quasi-experts were suitable stand-ins for experts, but only in the writing task, suggesting that expertise level and domain or task may, importantly, intersect.

Plucker, Kaufman, Temple, and Quian (2009) found that novices and expert critics' ratings of movies correlated very highly, although the experts gave substantially lower mean scores; agreement was also highest between critics and students who had the most exposure to movies, suggesting that cut-off points along the continuum of expertise can be hazy and affected by different types of experience. These equivocal results suggest that expertise in a field only sometimes increases consensus, and it is not clear when or why judge expertise is vital. It is therefore important to better understand what other factors may be contributing to any difference between the ratings of expert and novice judges.

Length and type of domain experience. From the creativity literature, when details of the experts' experience are given, it can vary widely but often refers to years of work in a field. For example, Yuan and Lee's 2014 study used experts with at least eight years of professional experience in product design, while Daly, Seifert, Yilmaz, and Gonzalez (2016) specify that their expert judges had at least three years of mechanical engineering product design coursework. Amabile's 1982 artist judges rating collages had at least four to five years of studio artwork experience. There may also be differences between those working mainly in more theoretical or practical areas of their field (e.g., academics or critics vs. practitioners) and how they might assess creativity differently. The question therefore

remains: what type and length of experience constitutes sufficient expertise to make someone an 'appropriate judge'? The answer to this is likely to be interwoven with the task which judges are required to assess.

Task Selection

In theory, the CAT can be used to assess outputs from a nearly endless variety of domains with a range of tasks. Some researchers choose familiar domain tasks (e.g., paintings for the visual arts, stories for literature), while others use more stylized 'creativity tests' meant to capture creativity within a laboratory setting. This raises the possibility that there may be some tasks which have no 'appropriate' expert to assess them, a circumstance where task and judge expertise could intersect to potentially affect CAT reliability. The use of domain experts to judge tasks with which experts only have a tenuous association may or may not be appropriate given the level or type of work they may be asked to rate. Under such conditions is the CAT to be limited to only traditional tasks and domains (e.g., poetry, collage)?

If limits are not required, then is training or giving guidance to judges for a given task permissible? For example, Friedrich and Mumford (2009) used psychology postgraduate student judges (not artistic experts) and achieved high inter-rater agreement in judging a figural laboratory creativity task. It is worth noting that in their study it was reported that there was some prior training, practice, and discussion amongst judges before ratings commenced.

It could be argued that training judges is at odds with the original theoretical framework of the CAT, which suggests judges only use their own subjective opinion. It has been suggested that in some circumstances "calibration" of nonexpert judges may be acceptable to align their ratings to experts' ratings (Hennessey, Amabile, & Mueller, 2011, p. 258-259). The importance, however, of theoretical neutrality and not biasing judges' subjective

opinions needs to be weighed against the appropriateness of particular judges for a particular type of task.

Additionally, a task can require the creation of a single artifact or require several from each participant; some tasks are simple, while others are complex. These task characteristics and their impact on rater fatigue are likely to be important for CAT reliability.

Number of Judges

In addition to judge experience, the number that are asked to act as judges may also be a factor for CAT reliability. Amabile (1982) stated it was difficult to specify an ideal number, only that it should be determined largely in terms of what was needed to achieve acceptable agreement. Still, it is difficult to plan what number is likely to achieve this aim ahead of time. As Silvia et al. (2008) point out, there is no definite guideline beyond vague intuition: “One is clearly not enough; 20 seems like overkill” (p. 81). Kaufman, Plucker, et al. (2008) have suggested that "for most purposes 5-10 judges is an adequate number" (p.74).

The number used in practice, however, varies greatly. In Amabile’s original multi-study CAT work in 1982, judge numbers ranged from three to 21. In practice, some researchers have used as few as two: Daly et al. (2016) asked two judges to rate the creativity of 439 engineering student designs, achieving an agreement of .70. Others have used much larger numbers of judges, especially when the judges’ ratings needed to be split into smaller experimental groups: e.g., Valgeirsdottir, Onarheim, and Gabrielsen's (2015) used 134 general public judges to rate the creativity of two mobile phone products, where the research design required subsequently dividing judges into four experimental conditions.

The question to consider within this wide range is to what extent it is possible, or desirable, to standardize the number of judges required for a CAT study. Differing research conditions and variables partly explain this variation, but not completely. Larger numbers are usually preferable for increased statistical power, but this ideal must be balanced against the

realistic feasibility of recruiting large numbers of experts to take part. It may also be possible to unintentionally inflate consensus through increasing the number of judges, as previously noted regarding the discussion of expert vs. novice judges in Kaufman, Baer, et al. (2008). Parallels, perhaps, can be drawn between the risks associated with a Type I and Type II error when exploring multiple correlations within a study. The implication of a Type I/Type II error equivalent for CAT protocol is that increasing the number of judges above 10 (the upper limit suggested by Kaufman, Plucker, et al., 2008) likely raises the risk of a Type I error (false positive) on inter-rater reliability. Conversely, reducing the number of judges below five increases the risk of both a Type I and Type II (false negative) error.

Stimuli Presentation

When embarking on a CAT study, depending on the task selected, decisions must be made about how best to provide or display the items to the judges. When numbers of items to be rated are small and the judges have access to the same location, providing physical objects to examine and handle may be relatively simple to organize. Often implementing the CAT is not this straightforward: e.g., expert judges are often busy, geographically disparate, and research design can require judges to rate a very large number of items, which takes time, can be tiring, and make comparative rating difficult or even invalid. In these circumstances, the method of presentation needs to be considered carefully, including the format or platform through which the items will be presented, how much judges will be able to interact with and manipulate the items they are viewing, the order of presentation, and the number of items the judges are asked to rate. New digital platforms could help solve some of these issues, but could potentially create new problems and would require some degree of usability testing (Barbot, Orriols, & Pouyade, 2008; Cseh, Jeffries, Lochrie, Egglestone, & Beattie, 2016).

Number of items and rater fatigue. Amabile's (1996) guidelines state that the CAT should be conducted in a relativistic fashion, i.e., comparing items within a sample to one

another, rather than on absolute or wider criteria. However, no known CAT studies have examined how many items judges are able to effectively compare to one another at a time in terms of relative creativity, although there is literature within cognitive psychology about choice overload and cognitive load (Miller, 1956; Scheibehenne, Greifeneder, & Todd, 2010). There is some suggestion that there is a link between time spent on the rating task and interjudge agreement from Amabile's own study (1982, Study 1). Findings inferred that the more data judges are asked to manage at a time and the longer they must spend on the task, the greater the fatigue, and the more the ability to compare relatively becomes compromised.

In the CAT literature, the number of items each judge rates shows a substantial range, from single figures to hundreds and possibly thousands. For example, Karwowski et al. (2016, Study 8) reused the data gathered by Jauk, Benedek, Dunst, and Neubauer (2013); both studies explored the relationship between creativity and intelligence. In the original data set, each judge would appear to have assessed 297 participants' creativity outputs with an average of 12 responses in total: i.e., 3,564 ratings per judge.

Ratings Procedure

Relative assessments. An aspect considered vital to the CAT procedure which is not necessarily noted in methodology reports is how judges are instructed to compare the works, and if they are explicitly told to make relative judgments within the context of the presented stimuli set, or on the basis of their absolute knowledge of the field in general (e.g., comparing an art student's work to other art students' work from the same study sample, as opposed to comparing it to the works of Picasso).

Factors rated. In its practical application, CAT guidance (Amabile, 1982, 1996) states that when creativity is assessed in a new domain (i.e., one that has not been studied with a particular task), judges should rate additional constructs – most commonly in artistic tasks,

technical execution and aesthetic appeal (Amabile, 1982; Hennessey, 1994) – to check whether creativity ratings are distinct from these criteria, for the sake of construct validity. Aside from technical execution and aesthetic appeal, researchers have rated, e.g., quality, originality, elaboration, or elegance, alongside creativity.

Within the CAT literature, some researchers have created instructions that directly ask judges to discount technical execution from their creativity rating (Baer, 1993). Some other researchers ask judges to rate creativity alongside technical execution, aesthetic appeal (e.g., Christiaans & Venselaar, 2005; Valgeirsdottir et al., 2015) or other factors; some only do this the first time they undertake a new CAT task (e.g., Hennessey, 1994; Kaufman, Baer, et al., 2008), while some do not distinguish between these and only measure one 'creativity' factor (see Jeffries et al., 2017, for review). Therefore, more work is warranted to better understand when creativity should or should not be clearly separated from other factors.

Rating scale. Amabile's early studies (1982) appeared to favor a rating scale of five points, though a variety of different scales were explored, including ranking, continuous scales, and categorizing into low, medium, and high. Preston and Colman (2000) indicate that an optimal rating scale should have a granularity of between five and seven points, and that reliability, validity, and discriminatory power suffer with rating scales that either have fewer or more decision points. There is also a longstanding debate about the pros and cons of even- vs. odd-numbered scales (Krosnick & Presser, 2010).

In CAT research, studies range between a three to ten point scale, and often very little or no justification is offered for why one scale was chosen over another. Kwon, Bromback, and Kudrowitz (2017) used a three-point scale with three expert judges to rate students' ideas for bicycle accessories, rating three factors (originality, feasibility, and marketability). Kwon et al. justified this by noting that "the three point system was used because the reviewers had to rate hundreds of ideas, and it makes it easier for them to categorize" (p. 2), highlighting the

need to balance ideals with realities, i.e., the oversimplicity of the scale and the possibility of reduced rater fatigue. Jauk et al. (2013) used a four-point scale in their study of creativity and intelligence. Kaufman, Evans, and Baer (2010) used a five-point scale to assess the visual, mathematical, verbal, and scientific creativity of school children. Kaufman et al. (2008) and Jeffries et al. (2017) each used six-point scales, though they asked participants to categorize items into three categories (low, medium, high creativity) first. Daly et al. (2016) used a seven-point scale. Harvey (2013) used a nine-point scale with three judges to measure the total creativity of a poster task for the 2012 Olympic games. Yuan and Lee (2014) used a 10-point scale with three judges to rate three variables: creativity, technical quality, and aesthetics.

Reliability, Agreement, and Statistical Test Choice

Kaufman, Plucker, et al. (2008) suggest the most frequent tests used to calculate inter-rater agreement are Cronbach's alpha, the Spearman-Brown prediction formula test, and intraclass correlation, and that they tend to produce similar results, offering the potential to be largely interchangeable. For practicalities of calculation, Cronbach's alpha has become the convention in most CAT studies.

What is actually inferred by ‘inter-rater agreement’, however, is worth considering, and a debate found in the CAT literature. For example, Stefanic and Randles (2015) noted a significant difference between internal consistency vs. absolute agreement – i.e., “if one judge’s creativity ratings are always two points higher than another judge’s ratings, then the two judges are consistent, but they would not be in agreement with what constitutes a given level of creativity” (p. 281). Shrout and Fleiss (1979) originally identified six different versions of intra-class correlation (ICC), later expanded to ten by McGraw and Wong (1996, as cited in Koo & Li, 2016), which depend on sampling method of judges as well as the items to be rated, single rater reliability vs. mean of multiple raters, and whether agreement is

relative or absolute. Different forms of ICC result in different interpretations and declaring the type of ICC used is therefore important. Koo and Li (2016) provide useful guidelines for selecting the appropriate ICC test for each situation, and how to report and interpret results. Both Shrout and Fleiss in 1979 and Koo and Li in 2016 note that many researchers fail to distinguish between these types of test.

Tests also depend on the type of data the rating scale has produced – continuous, ordinal, or nominal – as well as factors such as whether the data are normally distributed and if all the raters rated all the items or whether they each rated different subsets of the data (as in Cseh, 2014). Most CAT data are collected in the form of a Likert type scale with one creativity item per creative product. This brings up an unsettled debate about the type of data Likert type scale data in fact are - ranked ordinal (Jamieson, 2004) or continuous interval data (Pell, 2005), which has implications for which tests are the most appropriate to use with these data.

There are therefore reasons to reconsider whether the conventional Cronbach's alpha is the most appropriate test in all circumstances. It is considered by many to be the correct test for *internal* consistency on *continuous* data, while intraclass correlation can be used for test-retest and inter-rater reliability measurement on continuous data for *either* internal consistency or *absolute* agreement depending on design, while the considered-correct test coefficient for inter-rater reliability for normally-distributed, ordered rank data is weighted kappa, or Kendall's tau for nonparametric data (King's College London, 2017; Shrout & Fleiss, 1979; see Figure 1 for an example of how test decisions may depend on characteristics/goals of each particular study). However, see Myszkowski and Storme (this issue) for a critique of Cronbach's alpha and other traditional techniques of assessing the CAT, including the benefits of factor analysis/McDonald's omega, and tests taking judge characteristics into account.

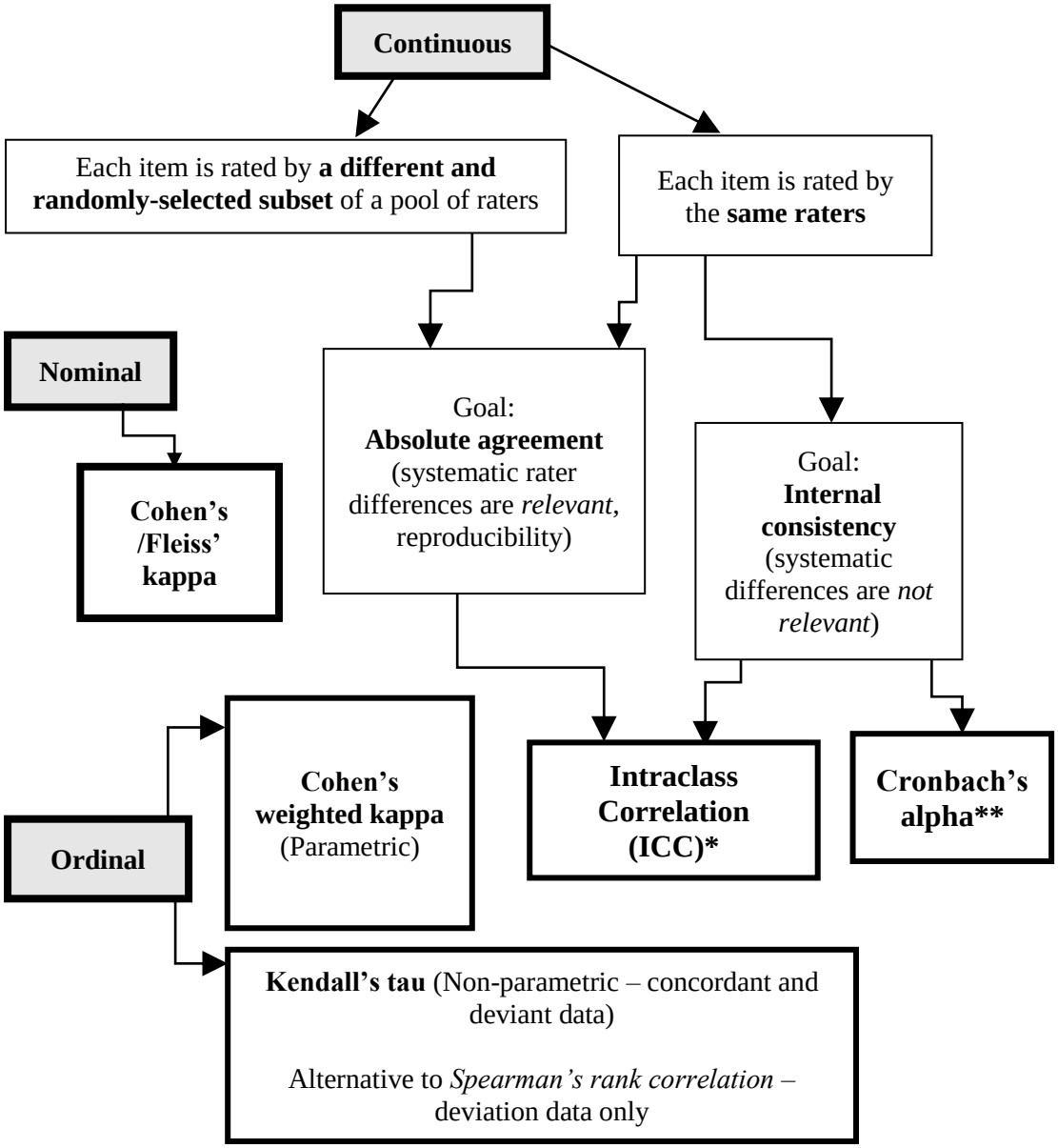

Figure 1. A decision tree illustrating which test of inter-rater reliability is traditionally considered appropriate under which circumstances (adapted from King's College London, 2017). *See Shrout and Fleiss (1979) and Koo and Li (2016) for discussion of the further complexities of ICC. **See Myszkowski and Storme (this issue) for a discussion of limitations and alternative test options.

Though the tests may seemingly produce roughly similar results, there is not always evidence that this has been checked, and the reasons for choosing one test over another, in the

absence of explicit justification, can seem arbitrary or based on convention alone. Likewise, the use of other measures to calculate interrater agreement (IRA) popular in other fields of study, such as Finn's r_{wg}, raises similar issues (O'Neill, 2017). While r_{wg} is not a measure commonly used in CAT studies, it has been used by some (e.g., Wigert, Reiter-Palmon, Kaufman, & Silvia, 2012).

Conclusions and Future Directions

Although the CAT is called a 'technique' that is credited to Amabile, the method, at its core, predates the 1982 operationalization and labeling of it. Notwithstanding a range of approaches similar to the CAT, even within CAT studies practice can vary considerably. The ultimate question is: how much *should* the CAT be a set protocol that researchers follow, and how much a set of loose principles that have acknowledged limitations and strengths adaptable to a given situation, domain, or design? It then follows, what parts of the CAT should be universal for scientific rigor, and which can be adaptable and organic so as to be applicable to each specific context?

The answer to these questions, ultimately, largely depends on who 'owns' or is most suitable to decide on the CAT protocol? As it has been over 35 years since Amabile first presented and defined it, and it has inevitably grown and changed over the years with other researchers amending and adapting the technique, it can be argued that the CAT could, now, belong to the creativity research community as a whole, and as such we must all take responsibility to ensure its quality control and evidence base for good practice.

The purpose of this article was to spark the needed debate and deeper investigation amongst the community of past, current, and future creativity researchers about the CAT's implementation. As illustrated here, there are a number of issues to settle about CAT procedure (see Table 1 for a summary). Such examples highlight the necessity for creativity researchers worldwide to work together to establish more solid, evidence-based standards of

consistency and transparency for the CAT procedure going forward, and to better understand the cumulative impact of even seemingly small variations in protocol if we are to have a better understanding of how creativity is assessed by social consensus.

Table 1

Summary of themes and questions arising from this review for future debate and research.

Themes	**Specific Questions**
Conceptual: The philosophical assumptions underpinning the CAT	• What distinguishes or uniquely characterizes the CAT, as opposed to other consensus measures of creativity? • Who, if anyone, 'owns' or 'defines' the CAT? To what extent should the CAT be a set protocol of operation to ensure scientific rigor, and to what extent is it a loose set of principles that can withstand adaptation? • How do we ensure integrity and quality control of CAT studies in the future?
Methodological: What is the link between procedural choices and outcomes?	• What level and type of expertise is appropriate for any particular given task? • Can judges be trained or given explicit definitions of creativity without losing the theoretical neutrality underpinning the CAT? Is theoretical neutrality vital in all circumstances? • How many items can judges reliably rate at once, relative to one another? • How does the way the material is presented affect ratings? • How does the granularity of the rating scale affect ratings?
Analytical: What statistical analysis techniques are most appropriate for each situation; what do they tell us?	• How many judges are required to establish reliable consensus? • Are Cronbach's alpha and ICC always the best 'conventions' for CAT inter-rater reliability calculations? • What kind of inter-rater reliability/agreement are we most interested in achieving?

There is a need especially for more experimental methodology studies, as well as systematic reviews and meta-analyses, so that a clearer picture can be formed of CAT research as it currently stands across the varied disciplines that use the CAT, and to measure the impact of different methodological choices on outcomes. This article offers a framework

for renewed discussion amongst researchers, perhaps at special CAT-dedicated symposia, or a new program of assessment methodology research, to ensure that the CAT remains one of the 'gold standards' of creativity research and assessment.

References

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. (1996). *Creativity in context: Update to the social psychology of creativity*. Boulder, CO: Westview Press.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, *16*(1), 113–117.

Baer, J., Kaufman, J. C., & Riggs, M. (2009). Rater-domain interactions in the Consensual Assessment Technique. *International Journal of Creativity and Problem Solving, 19*, 87-92.

Baer, J., & McKool, S. S. (2009). Assessing creativity using the Consensual Assessment Technique. In *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education* (pp. 1–13). https://doi.org/10.4018/978-1-60566-667-9.ch004

Barbot, B., Orriols, E., & Pouyade, H. (2008). Consensual assessment technique-interface (CAT-i). Copyright Cat-i.org.

Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, *132*(4), 355–429. https://doi.org/10.3200/MONO.132.4.355-430

Byrne, C., MacDonald, R., & Carlton, L. (2003). Assessing creativity in musical compositions: Flow as an assessment tool. *British Journal of Music Education*, *20*(3), 277–290. https://doi.org/10.1017/S0265051703005448

Christiaans, H. H. C. M., & Venselaar, K. (2005). Creativity in design engineering and the role of knowledge: Modelling the expert. *International Journal of Technology and*

Design Education, *15*(3), 217–236. https://doi.org/10.1007/s10798-004-1904-4

Cseh, G. M. (2014). *Flow in Visual Creativity* (Unpublished doctoral dissertation). University of Aberdeen, Scotland.

Cseh, G. M., Jeffries, K. K., Lochrie, M., Egglestone, P., & Beattie, A. A. (2016). A DigitalCAT: A fusion of creativity assessment theory and HCI. In *30th British Human-Computer Interaction Conference (Fusion Theme, WiP Track)*. (Bournemouth University) Poole, England. Retrieved from http://ewic.bcs.org/content/ConWebDoc/56962

Daly, S. R., Seifert, C. M., Yilmaz, S., & Gonzalez, R. (2016). Comparing ideation techniques for beginning designers. *Journal of Mechanical Design*, *138*(10), 101108.

Freeman, C., Son, J., & McRoberts, L. B. (2015). Comparison of novice and expert evaluations of apparel design illustrations using the consensual assessment technique. *International Journal of Fashion Design, Technology and Education*, *8*(2), 122–130. https://doi.org/10.1080/17543266.2015.1018960

Friedrich, T. L., & Mumford, M. D. (2009). The effects of conflicting information on creative thought: A source of performance improvements or decrements? *Creativity Research Journal*, *21*(2–3), 265–281. https://doi.org/10.1080/10400410902861430

Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*(9), 444–454.

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. A., & Mueller, J. S. (2011). Consensual assessment. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of Creativity* (2nd ed., Vol. 1, pp. 253–260). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-375038-9.00046-7

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38,* 1212–1218.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, *41*(4), 212–221. https://doi.org/10.1016/j.intell.2013.03.003

Jeffries, K. K., Zamenopoulos, T., & Green, A. J. K. (2017). Design creativity, technical execution and aesthetic appeal: A CAT with caveats (Part 2). *International Journal of Design Creativity and Innovation*, 6(1-2), 66-79. doi: 10.1080/21650349.2017.1381043

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

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the Consensual Assessment Technique. *Journal of Creative Behavior, 43,* 223-233.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton∗, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171–178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 332–340. https://doi.org/10.1037/a0034809

Kaufman, J. C., Evans, M. L., & Baer, J. (2010). The American Idol Effect: Are students good judges of their creativity across domains? *Empirical Studies of the Arts*, *28*(1), 3–17. https://doi.org/10.2190/EM.28.1.b

Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, *49*(3), 260–265.

https://doi.org/10.1177/001698620504900307

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. *Essentials of psychological assessment series*. Retrieved from http://www.loc.gov/catdir/enhancements/fy0827/2008008343-d.html%5Cnhttp://www.loc.gov/catdir/enhancements/fy0827/2008008343-t.html

King's College London. (2017). *Statistics advisory service FAQ - Questionnaire development.* Retrieved from https://www.kcl.ac.uk/ioppn/depts/BiostatisticsHealthInformatics/SAS/faqs9.aspx#a9_3

Kline, P. (2000). *The handbook of psychological testing* (2nd ed., Vol. 25). London, England: Routledge. https://doi.org/10.1016/S0191-8869(96)90047-1

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155-163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Bingley, England: Emerald Group Publishing Ltd. https://doi.org/10.1111/j.1432-1033.1976.tb10115.x

Kwon, J., Bromback, L., & Kudrowitz, B. (2017). Divergent thinking ability + interest = creative ideas: Exploring the relationship between cognitive creativity assessments and product design idea generation. In *Proceedings of the 29th International Conference on Design Theory and Methodology (DTM)* (pp. 1–5). Cleveland, OH: ASME. Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/185242/DETC2017-67261-0.pdf?sequence=1&isAllowed=y

McClary, R. B. (2009). *An investigation into the relationship between tolerance of ambiguity and creativity among military officers* (Unpublished doctoral dissertation). Kansas State

University, Manhattan, KS. Retrieved from http://krex.k-state.edu/dspace/bitstream/handle/2097/2210/RobMcClary2009.pdf?sequence=5

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8022966

Myszkowski, N., & Storme, M. (*this issue*). Judge response theory? A call to upgrade our psychometrical account of creativity judgements [Special issue]. *Psychology of Aesthetics, Creativity, and the Arts.*

O'Neill, T. A. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. *Frontiers in Psychology, 8,* Article 777. https://doi.org/10.3389/fpsyg.2017.00777

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education, 39,* 970.

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology and Marketing, 26,* 470-478.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

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

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research, 37*(3), 409–425. https://doi.org/10.1086/651235

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

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

Valgeirsdottir, D., Onarheim, B., & Gabrielsen, G. (2015). Product creativity assessment of innovations: Considering the creative process. *International Journal of Design Creativity and Innovation*, *3*(2), 95–106. https://doi.org/10.1080/21650349.2014.954626

Wigert, B., Reiter-Palmon, R., Kaufman, J. C., & Silvia, P. J. (2012). Perfectionism: The good, the bad, and the creative. *Journal of Research in Personality, 46*(6), 775-779.

Yuan, X., & Lee, J.-H. (2014). A quantitative approach for assessment of creativity in product design. *Advanced Engineering Informatics*, *28*(4), 528–541. https://doi.org/10.1016/j.aei.2014.07.007