AMERICAN PSYCHOLOGICAL ASSOCIATION DIVISION 10 Psychology of Aesthetics, Creativity, and the Arts

© 2019 American Psychological Association
1931-3896/19/$12.00

2019, Vol. 13, No. 2, 167–175
http://dx.doi.org/10.1037/aca0000225

Judge Response Theory? A Call to Upgrade Our Psychometrical Account of Creativity Judgments

Nils Myszkowski
Pace University

Martin Storme
Université Paris Descartes

The Consensual Assessment Technique (CAT)—more generally, using product creativity judgments—is a central and actively debated method to assess product and individual creativity. Despite a constant interest in strategies to improve its robustness, we argue that most psychometric investigations and scoring strategies for CAT data remain constrained by a flawed psychometrical framework. We first describe how our traditional statistical account of multiple judgments, which largely revolves around Cronbach's α and sum/average scores, poses conceptual and practical problems—such as misestimating the construct of interest, misestimating reliability and structural validity, underusing latent variable models, and reducing judge characteristics as a source of error—that are largely imputable to the influence of classical test theory. Then, we propose that the item–response theory framework, traditionally used for multi-item situations, be transposed to multiple-judge CAT situations in Judge Response Theory (JRT). After defining JRT, we present its multiple advantages, such as accounting for differences in individual judgment as a psychological process—rather than as random error—giving a more accurate account of the reliability and structural validity of CAT data and allowing the selection of complementary—not redundant—judges. The comparison of models and their availability in statistical packages are notably discussed as further directions.

Keywords: classical test theory, item–response theory, consensual assessment technique, creativity judgment, creativity assessment

Although various methods have been imagined to assess creativity, a substantial amount of research relies on Amabile's (1982) Consensual Assessment Technique (CAT), which consists of asking experts to evaluate creative products (Baer & McKool, 2009). Extensive research has provided a set of methodological guidelines on how to best collect accurate judgments of creative products. However, these methodological recommendations are often about how to better prepare (e.g., Storme, Myszkowski, Çelik, & Lubart, 2014) or select judges (e.g., Kaufman, Baer, Cole, & Sexton, 2008). In contrast, there have been much fewer investigations regarding how to examine the robustness of CAT data or how to obtain accurate composite scores for the measured attribute.

To examine the robustness of CAT data and derive composite scores for the attribute, researchers generally respectively compute Cronbach's α across judges and sum (or average) scores to aggregate judgments into a single score (Baer & McKool, 2009). There have been uses of latent variable models of judgment data (e.g., Myszkowski & Storme, 2017; Silvia et al., 2008) and discussions on how to investigate CAT data (e.g., Stefanic & Randles, 2015), but the general measurement framework to adopt to investigate the psychometric properties of CAT data and to obtain composite has not yet been discussed.

In this article, we discuss the typical psychometric investigations of CAT and creativity judgments, as well as describe the recurring challenges encountered. We trace them back to the underlying framework of Classical Test Theory (CTT) and subsequently present the framework of Item–Response Theory (IRT) as a more coherent and useful approach to CAT data.

The Limitations of Our Current Psychometrical Practice

While the CAT is an important advance in the measurement of product creativity, the employed statistical techniques that are commonly used, in both psychometric investigations and scoring strategies, result in critical challenges. In this section, we want to point to the main ones.

The Issues of Sum/Average Scoring

Typically, to aggregate the scores of judges in CAT and thus estimate a product's creativity—in other words, to achieve its measurement—researchers compute sums/averages across judg-

Nils Myszkowski, Department of Psychology, Pace University; Martin Storme, Laboratoire Adaptations Travail-Individu, Université Paris Descartes.

Correspondence concerning this article should be addressed to Nils Myszkowski, Department of Psychology, Pace University, Room 1315, 41 Park Row, New York, NY 10038. E-mail: nmyszkowski@pace.edu

ments. The underlying theory behind this practice is known as CTT or true score theory. At the core of CTT is the idea that an observed score is composed of a true score—fixed for a person/product—and a random error component. Because of this error component, researchers sum/average several observed scores, with the expectation that the errors will cancel themselves out. This provides an estimation of the true score, which is thus essentially the expectation of the observed score (Borsboom & Mellenbergh, 2002).

However, sum/average scores are not always good proxies for constructs (Borsboom, 2006). For example, different judges may have different judging abilities, yet all judges are given the same weight in the product's overall score. Moreover, different judges may use the rating scale differently, using more or less extreme values, yet the judges that produce more extreme scores will weigh more in the sum/average scores than the others—without necessarily being more accurate. Finally, judges may be more or less accurate in judging at different levels—some expert painters may, for example, be more accurate judges of professional paintings than children's drawings—yet, these differences are not accounted for through sum/average scoring.

Additionally, in CTT, the person measurement is affected by the instrument's characteristics, and the instrument characteristics are affected by the person being measured (de Ayala, 2013), thus limiting the interpretation of both scores and judge characteristics. For example, the estimated attribute of a product through sum/average scores depends on the severity of the judges who judged it, while the properties of judges studied through the average score given by a judge depend on the attribute of the products judged.

The Conceptual Inconsistencies of CTT

From our previous description of the CAT, it appears that there are two causes of the set of observed scores: the product and the judge. We could summarize this situation through a model that describes an observed score as resulting from both the product's attribute and the judge's characteristics. While product and judge attributes (such as severity and discrimination) can be studied through average scoring and by using factor-analytic models—which, as we later argue, are closer to IRT than CTT (Mellenbergh, 1994)—these elements do not actually appear in the formulation of CTT (Borsboom, 2006), which creates a conceptual inconsistency between the situation and the statistical approach.

Moreover, the CTT is unfalsifiable—the fact that a score is equal to its expectation (the true score) and an error component is necessarily true, since both the true score and the error part are unknown (de Ayala, 2013)—which disqualifies it as a measurement model. This implies that, in CTT, researchers cannot test measurement models against data (Borsboom, 2006), which presents a conceptual problem when formulating hypotheses about the relations between the product and judge attributes and the observations.

The Routine Uses of Cronbach's α

In most domains of psychology, computing α has become a routine procedure to examine the psychometrical robustness of an instrument (McNeish, 2018). It has been advanced that it is due to it being more easily accessible than other methods (Borsboom, 2006), as well as to misconceptions regarding its use (Sijtsma, 2009). Notably, it was pointed out that researchers often incorrectly stretch the meaning of α, from a measure of internal consistency based on strict assumptions (which we later describe) to other qualities, such as reliability in general, or the degree of unidimensionality of an instrument (Sijtsma, 2009). In CAT research, this also applies, as (1) α—sometimes in addition to conceptually similar measures, such as intraclass correlation coefficients (Stefanic & Randles, 2015) or the Spearman-Brown corrected α (Kaufman et al., 2008)—is largely reported and discussed as the main measure of interrater reliability, and (2) it is implied that, because a set of judgments presents a high α—for which assumptions have generally not been checked or discussed—the judges necessarily assessed the same attribute of the product, which is the only explanation for the judges' scores (meaning that there is no other common factor or correlated errors between judges). As a consequence, creativity researchers often use α as a criterion that indicates judge expertise or content objectivity (Storme et al., 2014).

The Disregard for α's Assumptions and Biases

Essential τ-equivalence. Alpha relies on the assumption that the observations—here the judgments—should be replications of an identical measurement process with (potentially different) precision. In other words, α is based on the assumption of (essential) τ-equivalence, which states that the true score is constant—while the error may vary—across items/judges (Raykov, 1997).

In the context of CAT, the assumption of τ-equivalence implies that all judges rate products using the same scale (Graham, 2006). It means, for example, that a 1-point difference in the response scale necessarily holds the same meaning for all judges. The violation of this assumption implies misestimating reliability when using Cronbach's α. This is problematic because (1) this assumption is rarely tested, and (2) alternative estimates of reliability that do not formulate this assumption—such as Raykov's congeneric measure of reliability (Raykov, 1997)—are rarely used.

Unidimensionality. Alpha also assumes that a single attribute underlies the relations between the item scores. In other words, α relies on the assumption that the measure is unidimensional and presents only one common factor. Unidimensional measures, provided sufficient items, often (but not necessarily) tend to have high αs, but high αs can also be observed with multidimensional measures or when errors are correlated. Yet, researchers rarely test unidimensionality prior to reporting α. In fact, they often use α as a measure of both reliability and unidimensionality (Sijtsma, 2009).

The Absence of Conditional Reliability

Despite contemporary efforts (e.g., Raju, Price, Oshima, & Nering, 2007), in CTT, reliability is mainly conceptualized as a fixed property for a set of observations from an instrument. In other words, different observations have the same expected reliability. Thus, Cronbach's α is fixed across a sample. In CAT, it means that one cannot tell if a product's attribute has been more reliably measured than another's. It is problematic, as it would be logical to expect that different products with different levels of the attribute could be judged with different accuracy.

The Underuse of Generalizability Theory

Generalizability theory (Brennan, 1992) proposes an extension to CTT that integrates components of the observed score—typically, the grand mean, the person effect, the item effect, the rater effect, and the measurement occasion. Beyond offering a more extensive account of the variability of observed scores, generalizability theory is also interested in how measuring the reliability of scores depends on their intended use. For example, it distinguishes between reliability indices intended for relative decisions (e.g., to study a product's creativity as opposed to the other products in the sample) and reliability indices intended for absolute decisions (e.g., to qualify the individual who created a product as showing creative giftedness). In creativity research, with a few exceptions (e.g., Silvia et al., 2008), generalizability theory is certainly underused, as the decisions relative to the examination of reliability are rarely put in relation with the intended interpretation of the score.

The Underuse of Factor Analysis

Even though Cronbach's α is a heavily used reliability measure in CAT, researchers have explored more accurate methods to investigate the structure of the relations between the judgments of creative products—notably factor analysis. However, as we will discuss, it is often used with limited aims.

Its relevance. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) have a special place in psychometric research, in that they are close to CTT applications like Cronbach's α computationally—for the reason that they can be computed over a correlation or covariance matrix and do not require the full information in the data—while in reality they are *conceptually* similar to IRT, because of their latent variable formulation (Mellenbergh, 1994). More specifically, applied to CAT, factor analysis formulates that observed judgment scores are explained by latent product attributes and judge characteristics—which is in line with the definition of Judge Response Theory (JRT) that we later provide.

In addition, as unidimensionality is an assumption of α, EFA/CFA may be used to investigate dimensionality prior to computing α. Both EFA and CFA can be used in the study of unidimensionality: EFA allows one to explore potential multidimensional structures and potentially discard them, while CFA can offer model fit indices for a specified unidimensional structure.

Its underuse. The habitual use of traditional EFA and CFA is, however, often still problematic for various reasons. First, CAT data are mostly ordinal, and although there are methods for factor analysis models that can be applied with ordinal manifest variables (Li, 2016), they are largely underused. Thus, upon using (linear) EFA/CFA models, an interval level of measurement is typically assumed without test or discussion, often regardless of the number of Likert scale points and possible bound effects. Second, while factor analysis permits the study of characteristics of items/judges—such as discrimination (through loadings) and severity (through means)—these characteristics are in practice rarely discussed, and EFA/CFA is typically only used to discuss the number of latent attributes that underlie the data. Finally, and more importantly, EFA/CFA models are able to yield estimates of the latent attribute: the *factor scores.* Yet, factor analysis is often used as a verification of the dimensionality of an instrument, before reverting to sum/average scores (and computing α). This practice is inconsistent: If one concludes that a measurement model is appropriate, why not use the estimates of this very model to achieve measurement?

JRT

What Is IRT?

A critical advance in psychometric theory is that measurement does not consist of *replacing* a construct with a response but instead in *linking* the response to the construct (Borsboom, 2006). This advance has notably prompted the creation of the framework known as IRT, which essentially models item responses as a function of a case-dependent latent attribute—named θ—and item dependent characteristics.

The IRT framework is often argued to be *conceptually stronger* than CTT (Borsboom, 2006; Borsboom & Mellenbergh, 2002), in that it provides an actual model of the psychological situation of responding. Perhaps a greater difference between CTT and IRT is that CTT assumes the existence of a true score (de Ayala, 2013)—the expectation of the observed scores with an infinite number of items—while IRT assumes a semantic object—a *construct*—to have a causal effect on the observations (Borsboom, 2006; Borsboom & Mellenbergh, 2002).

Ordinal IRT Models

IRT is mainly known for its account of situations where linear approximations were evidently deficient, notably when modeling binary pass/fail item responses (de Ayala, 2013). Consequently, when IRT appears in psychological testing education, it is often discussed as a framework appropriate for binary items primarily (e.g., DeVellis, 2016). Yet, there is actually a considerable amount of work on the use of IRT for other responses scales. Regarding ordinal responses, a wealth of available models have been developed (see Thissen & Steinberg, 1986, for a taxonomy)—notably the Graded Response Model (Samejima, 1969), the Modified Graded Response Model (Muraki, 1990), the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982), and the Generalized Partial Credit Model (Muraki, 1992). Beyond models with two facets (items and persons or, in CAT, judges and products), these response models may be extended to models with more facets—for example, through the Many-Facet Rasch Model (Barbot, Tan, Randi, Santa-Donato, & Grigorenko, 2012; Linacre, Engelhard, Tatum, & Myford, 1994; also see Primi et al., 2019).

Introducing JRT

In CAT, when investigating reliability as well as in scoring, we essentially consider judges as items (Kaufman, Lee, Baer, & Lee, 2007) and proceed with our analyses and scoring in the same fashion as we would with any multi-item Likert-type instrument. From this item-to-judge translation, we propose that the IRT framework be applied to product judgments—a translation that we could possibly name JRT. Indeed, one possible reason for the underuse of IRT by creativity researchers might be that IRT *only* has something to do with items and that it is primarily meant for educational measurement. Thus, through this (perhaps unnecessary) reformulation, and through the applications that we later describe, we hope to encourage creativity researchers to consider the relevance of IRT in creativity research.

We could define JRT as a psychometrical framework that uses latent attributes—trait(s) and/or class(es)—of a stimulus or product and of a judge as predictors of observed judgments. For example, in a typical CAT application, the latent attribute of the product would be creativity, while the latent attributes of judges would be severity and accuracy.

The Benefits of JRT

Just as the IRT framework presents multiple advantages over CTT, JRT can be beneficial for creativity researchers who use CAT. Throughout this section, we report proof-of-concept graphical outputs to demonstrate a few of the possibilities offered by JRT. To introduce further concepts and outputs, we simulated 5-point ordinal judgments of 30 products by four judges—which mimics a rather minimal CAT situation with only a few expert judges available to judge a limited number of products. The generation of the data was achieved with the package “mirt” (Chalmers, 2012) for R, based on a Graded Response Model. The parameters were chosen arbitrarily, with the only aim to make the plots later presented readable and conveniently discussed.

We fitted the responses with a Graded Response Model in “mirt,” and the plots were computed with the package “jrt” (Myszkowski, 2019) for R, currently under development. Again, this simulation was only meant as a proof of concept, in order to provide a snapshot of the possibilities offered.

JRT renders judge “brainwashing” useless. CTT renders desirable the repetition of parallel (or at least essentially τ-equivalent) judgments/items and leaves to the researchers’ control the balance between redundancy and inconsistency. This translates into paradoxical recommendations to scale developers, such as maximizing consistency but not content similarity (e.g., DeVellis, 2016). Similar to how researchers tend to build scales with redundant items to maximize consistency, when instructing judges, creativity researchers regularly overload judges with unnecessary instructions on how to use the response scale, so as to make observations more parallel. In JRT, however, differences between judges are accounted for by the model, which means that attempting to make judges form parallel judgments is unnecessary. Thus, in using JRT for creativity measurement, the instructions or training administered to judges can be solely focused on the *content validity* of what is judged, rather than on a specific expected use of the response scale.

JRT advances the study and account of judge variability. Because in JRT judges are allowed to have unique response functions, JRT is an ideal framework for *judge analysis*. The item response function plot—which presents the predicted probability of the different responses as a function of the latent trait θ—is regularly used to describe item functioning in IRT. Likewise, we show judge (category) response functions plots of multiple judges simultaneously in Figure 1. We present the model-estimated probabilities of choosing each response category (1, 2, 3, 4, and 5)—the vertical axis—as a function of varying levels of the latent trait θ—the horizontal axis. θ is typically (but not necessarily) assumed of a standard normal distribution: Thus, θ estimates can be interpreted as *z*-scores. Such plots allow seeing each judge’s use of the scale, as well as each judge’s thresholds between response points—for example, here we can see that Judges 3 and 4 did not use the maximal point or that average products are likely to receive scores of 3 by Judge 1 but scores of 2 by Judge 3.

JRT allows one to study judge variability. In contrast to CTT, in the IRT framework, since items have different ways of functioning (represented by item parameters), it is possible to study how different variables influence item functioning. This specific range of applications of IRT modeling is called differential item functioning in IRT—a nonlinear (or generalized) variant of studying measurement invariance in the linear factor-analytic tradition. Likewise, creativity researchers could take advantage of differential *judge* functioning to study how different characteristics of judges—expertise, training, or personality traits, for example—or of the product can influence judgment, or to account for such effects in our scoring strategies.

θ scores are *construct* estimates. When using average or sum scores, we try to estimate the true score—which is an expectation of the observed score, provided an infinite number of judges—but not a *construct* (Borsboom & Mellenbergh, 2002). In contrast, JRT directly estimates the latent trait (or class) estimate—θ—which is already standardized (in the case that a latent trait is considered) and may be directly used in further analysis. Thus, JRT is a coherent approach for both psychometric investigation and to estimate a construct for further analysis.

JRT allows one to study conditional reliability. In CTT, reliability is mostly conceptualized as a group-level estimate. In other words, within a group, all observed scores have the same expected reliability—and thus standard error of measurement. In contrast, in IRT, reliability is *conditional* upon the person/product attribute. In other words, in IRT, each observation has a different reliability estimate. This presents a direct advantage, as JRT makes it possible to take into account that judges could be more or less accurate at different levels of product creativity (or, in general, the attribute being measured). This implies that the question of the adequacy between judge severity and product creativity levels finds a direct statistical echo with JRT.

Reliability—which is related to information and standard errors—can be presented graphically through item information function plots. Taking the first judge as an example, we present a judge information function and judge reliability function plot in Figure 2.

In Figure 3, information and reliability functions are presented but with estimates of information and reliability that are marginalized across the entire set of four judges. IRT-based marginal reliability estimates are often used rules of thumb for acceptability that are similar to CTT-based estimates (e.g., Myszkowski & Storme, 2018)—even though decisions on acceptability should remain context dependent.

Related to this possibility, not only could we study conditional reliability when we study the psychometrical robustness of our creativity judgments, but we could also use conditional standard errors to weigh observations in statistical modeling or to filter out unreliably judged products from further analysis.

JRT provides model-based marginal reliability estimates. Although conditional reliability is a strength of IRT, one can also compute estimates of reliability that are marginalized across a distribution of θ estimates. Indeed, there are different situations where a marginal or group-level estimate can be desirable in JRT, such as when one wants to compare the overall reliability across groups of judges. There are different ways to achieve this in

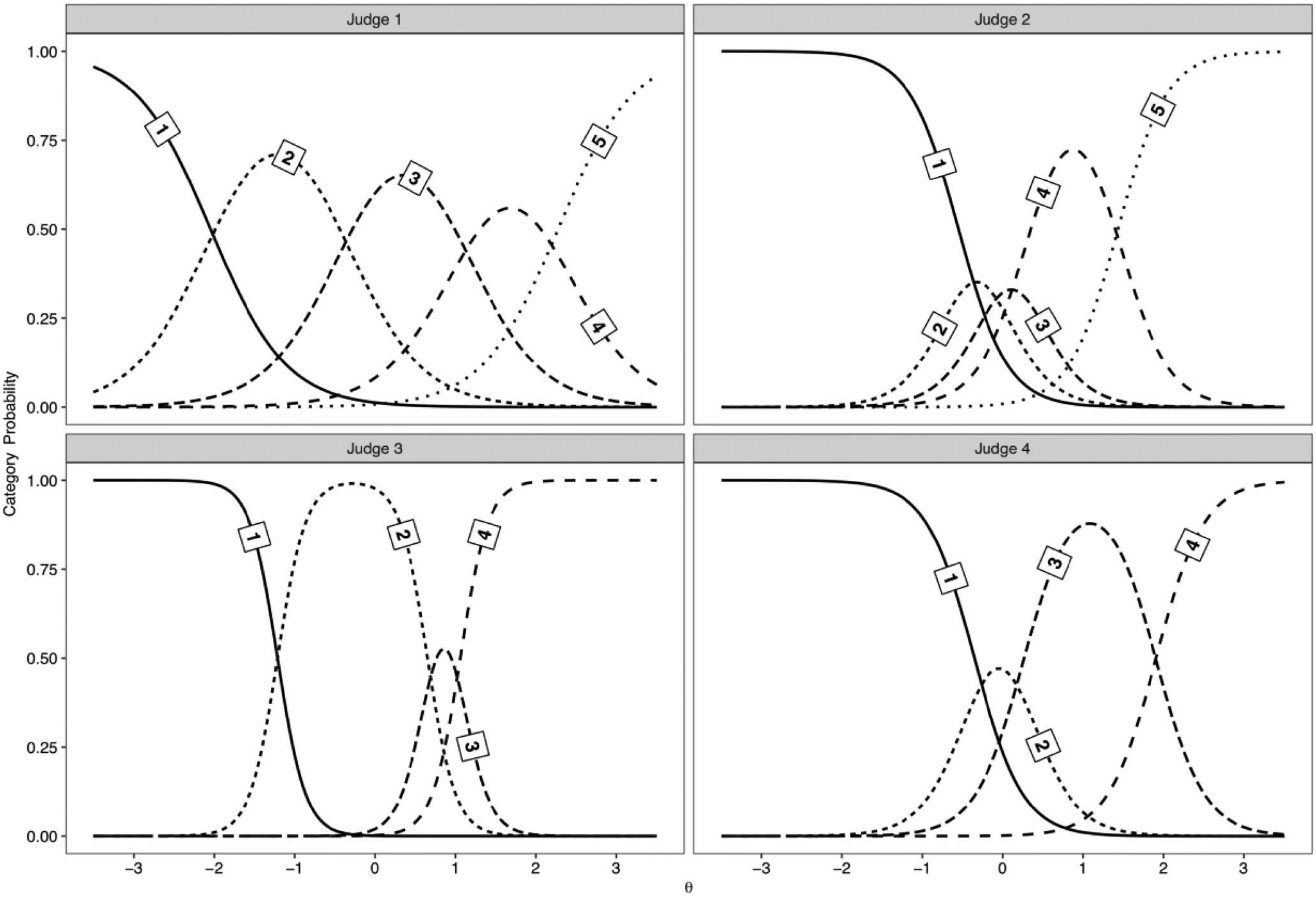

Figure 1. Example judge category response function plots.

IRT/JRT. Notably, an estimate of marginal reliability can be obtained through averaging the reliability across the observations of the sample—this is often called *empirical reliability*. An alternative strategy is to average reliability from an assumed prior density function (usually normal) of the attribute. Of course, both may be reported (e.g., Myszkowski & Storme, 2017, 2018). In addition, bootstrapping strategies may be used to draw inference on these reliability estimates (e.g., Myszkowski & Storme, 2018).

JRT allows one to explore dimensionality and test structural validity. As we previously noted, CTT identifies an unobserved true score as the sum of the observed score and an unobserved error, which is irrefutable. In contrast, the IRT/JRT framework offers models—where a set of person/product attributes and item/judge characteristics are theorized as explanations for the observed judgments—that are testable (Borsboom & Mellenbergh, 2002).

Although IRT and factor analysis are similar conceptually (Mellenbergh, 1994), they have different traditions regarding the procedures used to investigate model fit. The factor-analytic tradition is often concerned with exploring and concluding on the number of latent attributes underlying the data, notably using loadings and variance explained for EFA, as well as absolute model fit indices—such as the comparative fit index or the standardized root mean residual—for CFA. In contrast, the IRT tradition is to examine misfitting items (through item–fit statistics) and to compare response models with one another, often using likelihood ratio tests.

In addition, IRT research has made advances that allow its use to explore and test the dimensionality of an instrument—with tools that are similar to those of traditional factor analysis. Packages like “mirt” (Chalmers, 2012), for example, allow one to estimate altogether both exploratory and confirmatory models and to obtain indices known to researchers familiar with traditional EFA, such as factor loadings and absolute model fit indices. Therefore, JRT allows one to explore dimensionality and to compare models that may vary in both the number of judge characteristics to consider and the number of latent attributes of the products.

JRT is (potentially) economical. As we previously demonstrated, in JRT, each judgment of a product by a judge has its own reliability. Consequently, after an initial calibration of the model, the latent attribute and its reliability can be reestimated after each judgment. Thus, the current estimate of the attribute can be used to select the most appropriate next judge—the one that maximizes the expected gain in reliability. In addition, the current estimate of reliability may be used as a rule for stopping judgment—if the attribute is already reliably measured, it would not need to be judged again. That continuous estimation and intelligent item presentation process is known as computer-adaptive testing (Weiss, 1982) and is used with IRT to minimize the burden of

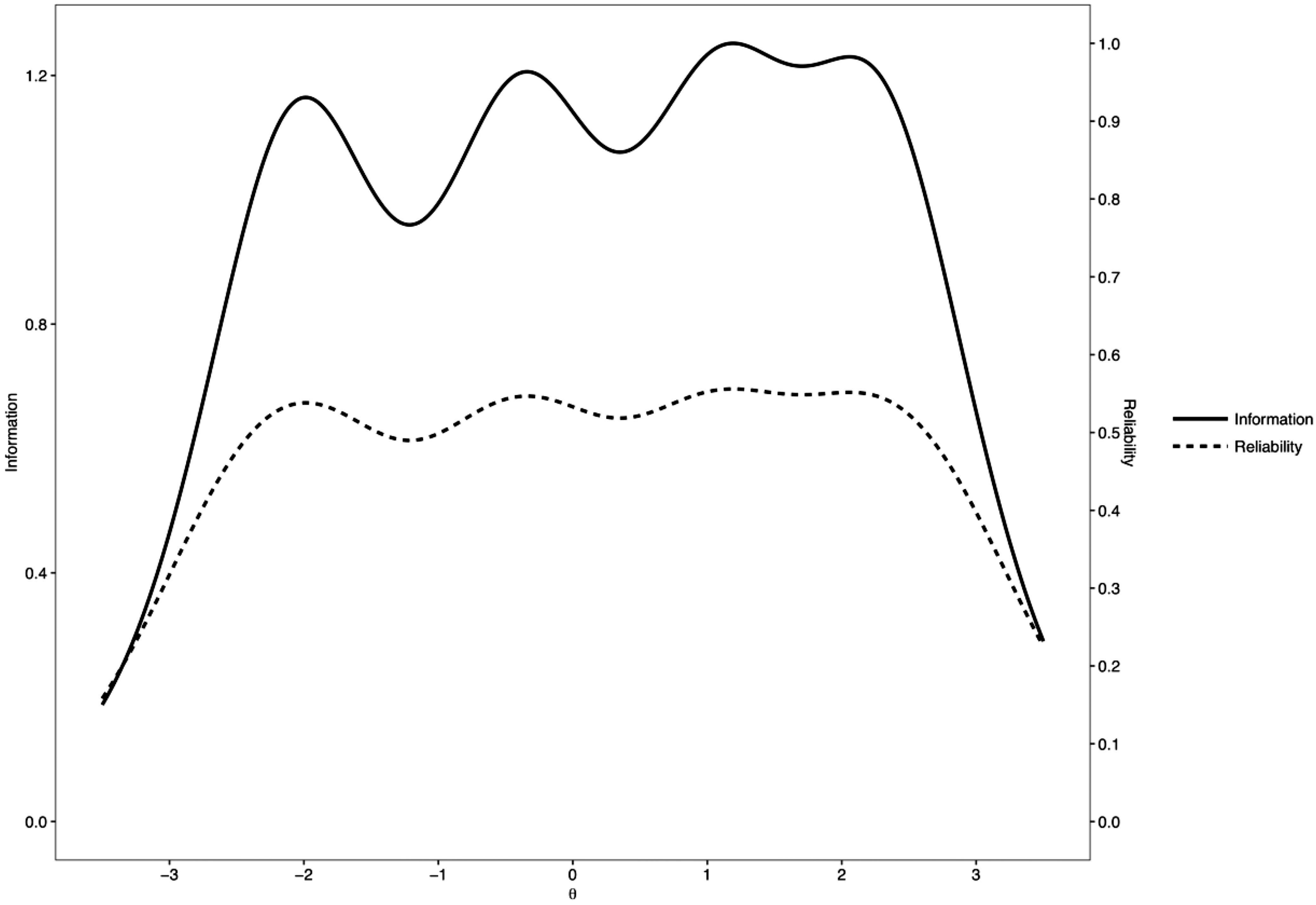

Figure 2. Judge information function and judge reliability function plot.

taking a large number of items while obtaining highly reliable scores.

Similarly, we could propose a computer-adaptive judging process, where the products would be channeled to only the most appropriate judges as a function of the product's current estimated creativity and where the products would stop being judged when their estimated creativity levels are reliable enough. To maximize feasibility, we could, for example, imagine contacting easily reachable novice judges in the initial estimation stage and then channeling products to be judged by appropriate expert judges. So far, such applications still appear impractical, but future research may come up with realistic ways to adapt this to CAT.

JRT allows one to combine different response scales together. In CTT, all the responses are assumed to use the same response scale, which constrains researchers to using only one response scale. What if researchers intended to use a combination of response scales? For example, what if researchers wanted some judges to judge "quickly" with a binary creative/noncreative scale and then some other judges to judge more extensively with an ordinal or a visual analog scale? Or, what if researchers collected data from multiple sources that already used different rating scales?

In JRT, because the construct and the observations are clearly distinguished, the same construct can easily be linked to (and thus measured by) different response scales at the same time. One could imagine applications such as measuring creativity through a combination of the judged creativity of a product, the time that was spent judging it, jointly measured. In addition, the IRT/JRT framework may be used to account for measurement situations that involve a multiplicity of measurement facets.

Discussion

The CAT (Amabile, 1982) is a central research methodology in creativity research, and its methods are actively debated. Yet, its current psychometrical treatment is mostly constrained to CTT-based methods—as shown by the extensive use of sum/average scoring and Cronbach's α—which is certainly approachable from a computational point of view (Borsboom, 2006) but severely limits our use and understanding of judgment data.

We proposed to consider the psychometrical treatment of CAT through JRT—a translation of the IRT framework to multiple-judge situations—as an alternative to CTT and presented its multiple benefits. While CTT reduces judge characteristics to random error and tries to eliminate it through summing/averaging, JRT includes judge characteristics in a measurement model. In doing so, JRT offers models consistent with the testing situation, which can be used for investigation and scoring. Furthermore, JRT opens

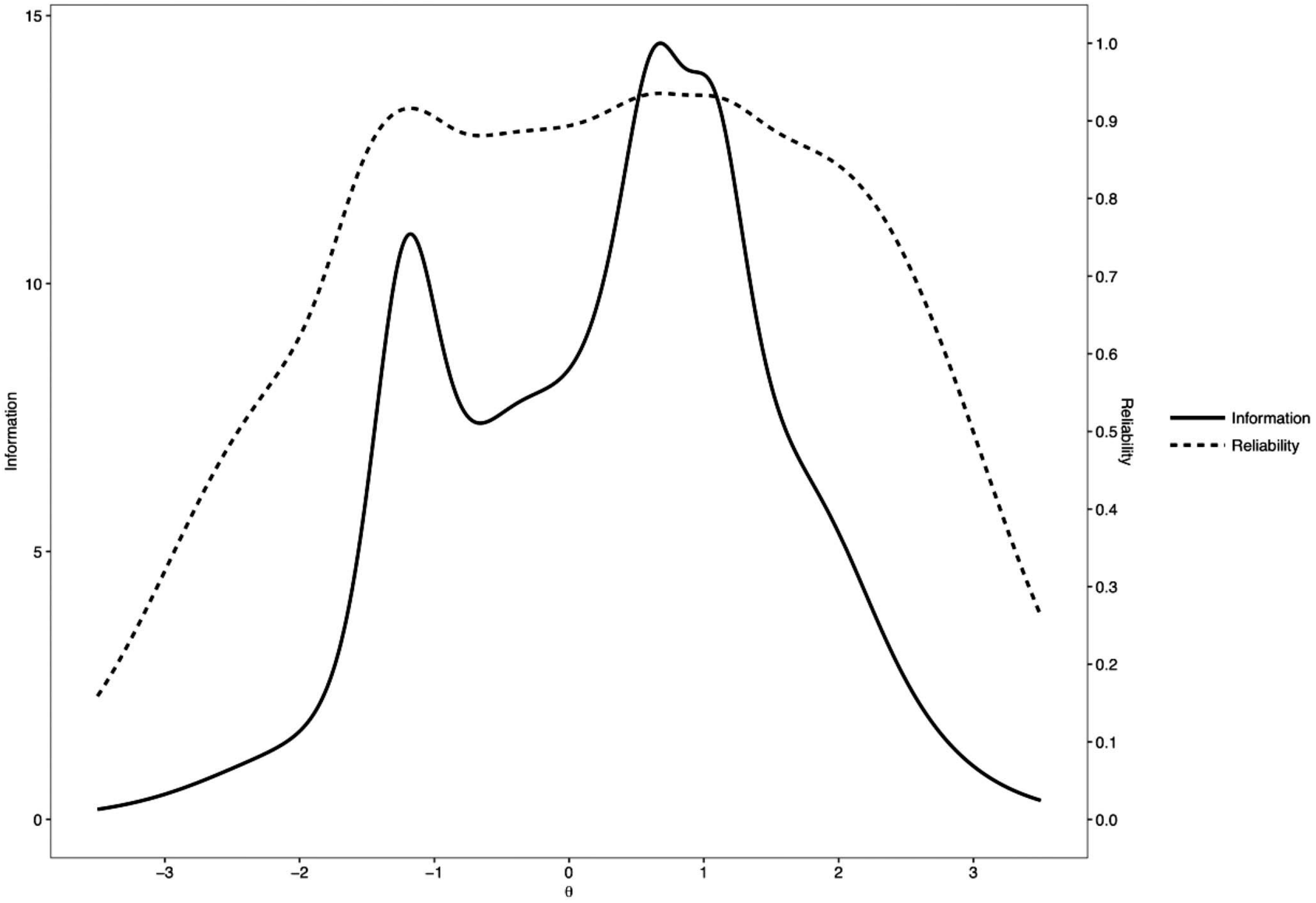

Figure 3. Total information function and total reliability function plot.

new doors for CAT research, notably by advancing the study of the variability of creativity judgments, as well as the understanding of reliability as conditional upon latent creativity levels, on the combination between responses of various response scales together, and on the optimization of judge selection procedures.

Further Directions

We focused on the advantages of the JRT framework and its appropriateness for CAT, but we did not discuss other connected important points that would need further reflection. First, we did not discuss the differences between the ordinal models as they apply to CAT. We focused on the benefits of the IRT/JRT framework, but of course, researchers who are interested in applying it to CAT judgments should be attentive to the variety of available models and what their differences are (see de Ayala, 2013, for an introduction). We also recommend empirical and theoretical model comparisons in the context of the CAT.

Second, it should be noted that we did not oppose the traditional IRT practice (which stems from logistic models) to the traditional practice of factor analysis (based on linear models). In fact, both continuous and ordinal models actually exist in both the IRT and the factor analysis tradition (Mellenbergh, 1994), and thus, our point here is that measurement models in general—which essentially link latent attributes to observations—should be considered as a more conceptually sound alternative to the CTT framework in scoring and psychometric investigations of CAT data.

Third, we did not discuss the various requirements of IRT/JRT in terms of sample size or missing data. This is because there are many different conditions (notably dimensionality, model complexity, judge characteristics, or estimation procedures) that impact sample size requirements and prevent from providing magic threshold numbers (de Ayala, 2013). Nevertheless, sample size is often considered the main obstacle in using IRT. A few important points must be made here. First, sample size requirements may be overestimated by researchers inexperienced in IRT, as recent advances in estimation procedures have reduced sample size needs (Zickar & Broadfoot, 2009), and this may not be reflected in psychometric trainings. Second, the object of study is an important consideration: Simulation studies of ordinal IRT models like the Graded Response Model indicate that the estimation of creativity levels would be especially impacted by a low number of judges, while the estimation of judge parameters would be primarily impacted by fewer judged products (Kieftenbeld & Natesan, 2012). For example, when using a Graded Response Model with marginal maximum likelihood estimation, as few as five items (judges) could already provide ability (creativity) estimates that are on average correlated at .84 with true attribute levels—with a negligible effect of how many products are judged (Kieftenbeld & Natesan, 2012). Thus, for purposes of measuring the product attribute, we could tentatively advance that as few as five judges may in some cases be sufficient, even though more research—on actual and simulated data—is needed. It is finally important to note that CTT and IRT are not comparable approaches in sample size requirements, as IRT formulates testable measurement models, while CTT is necessarily true with any or even no data (de Ayala, 2013). In other words, without using measurement models (from the IRT or the factor analysis tradition) to verify that sum/average scores are accurate proxies of accurately estimated constructs—by, for example, checking that sum scores correlate with attribute estimates of a well-fitting measurement model (e.g., Myszkowski & Storme, 2017)—CTT essentially avoids by assumption the issue that IRT attempts to solve.

Fourth, a major issue of the application of IRT is the limited availability of its methods in the most used statistical packages (Borsboom, 2006). Still, IRT modeling is available in several statistical packages, both software (e.g., StataCorp, 2017) and freeware (e.g., Chalmers, 2012). For researchers inexperienced in IRT, an easy way to start applying IRT to their CAT data with minimal coding expertise or data preparation work is to use the application "IRTShiny" (Hamilton & Mizumoto, 2017), which can be called locally from R or on a remote server. Despite not being fully customizable, "IRTShiny" allows one to fit popular ordinal models such as the Graded Response Model and Generalized Partial Credit Model.

Finally, this article mainly focused on a comparison between the IRT/JRT approach and the most common practice, which is use of sum scoring and Cronbach's α. It is, however, important to acknowledge that this does not represent research practice in its entirety. For example, some creativity researchers (e.g., Stefanic & Randles, 2015) have used intraclass correlation coefficients instead of α as a measure of reliability. In addition, the fields of psychology of creativity and empirical aesthetics have previously made uses of IRT—or of its extensions—to model judgments (e.g., Barbot et al., 2012; Myszkowski & Storme, 2017; Silvia, Martin, & Nusbaum, 2009; Tan et al., 2015). However, IRT models remain underused, and as we mentioned, the question of the psychometrical framework per se had not been debated in creativity measurement.

Conclusion

Creativity researchers are traditionally innovative in their research methods, ready to face statistical challenge, and prompt to overcome methodological dogma. Yet, like most psychology researchers, they may overlook what many psychometricians discussed as nothing more than a silent revolution in psychological measurement (Cliff, 1992).

IRT/JRT certainly presents its own set of challenges, and therefore, we would not recommend abandoning CTT in all cases, as it still provides a set of tools that are undoubtedly practical, with assumptions that in some cases can be reasonable or necessary. Nevertheless, we recommend that CAT researchers refocus their efforts on the measurement situation and how to best represent it with an actual measurement model. In doing so, they could be able to achieve both a more accurate measurement of product attributes and a better understanding of the judgment process. For this reason, we recommend that IRT/JRT be considered as an alternative to sum/average scoring and CTT-based reliability estimates such as Cronbach's α.

The psychometrical methods used in creativity research probably do not lag behind those of other fields, and our aim is not to diminish or criticize major advances in CAT and in creativity research. Instead, we hope to encourage creativity researchers to challenge their training and habits by shedding light on the numerous conceptual and practical benefits of a (roughly) *novel* and (promisingly) *useful* psychometrical framework.

References

Amabile, T. M. (1982). *The social psychology of creativity: A consensual assessment technique*. Retrieved from http://www.hbs.edu/faculty/Pages/item.aspx?num=7355

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573. http://dx.doi.org/10.1007/BF02293814

Baer, J., & McKool, S. S. (2009). Assessing Creativity Using the Consensual Assessment Technique. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). Hershey, PA: IGI Global. http://dx.doi.org/10.4018/978-1-60566-667-9.ch004

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

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. http://dx.doi.org/10.1007/s11336-006-1447-6

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30,* 505–514. http://dx.doi.org/10.1016/S0160-2896(02)00082-X

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11,* 27–34. http://dx.doi.org/10.1111/j.1745-3992.1992.tb00260.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186–190. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00024.x

de Ayala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford.

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66,* 930–944. http://dx.doi.org/10.1177/0013164406288165

Hamilton, W. K., & Mizumoto, A. (2017). *IRTShiny: Item response theory via Shiny* (Version 1.2). Retrieved from https://CRAN.R-project.org/package=IRTShiny

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,* 171–178. http://dx.doi.org/10.1080/10400410802059929

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2,* 96–106. http://dx.doi.org/10.1016/j.tsc.2007.04.002

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36,* 399–419. http://dx.doi.org/10.1177/0146621612446170

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21,* 369–387. http://dx.doi.org/10.1037/met0000093

Linacre, J. M., Engelhard, G., Jr., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21,* 569–577. http://dx.doi.org/10.1016/0883-0355(94)90011-6

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. http://dx.doi.org/10.1007/BF02296272

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23,* 412–433. http://dx.doi.org/10.1037/met0000144

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307. http://dx.doi.org/10.1037/0033-2909.115.2.300

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71. http://dx.doi.org/10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an Em algorithm. *ETS Research Report Series, 1992,* i–30. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01436.x

Myszkowski, N. (2019). jrt: Item Response Theory Modeling and Scoring for Judgment Data (Version 1.0.0) [Computer software]. Retrieved from https://CRAN.R-project.org/package=jrt

Myszkowski, N., & Storme, M. (2017). Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). *Personality and Individual Differences, 117,* 91–100. http://dx.doi.org/10.1016/j.paid.2017.05.041

Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence, 68,* 109–116. http://dx.doi.org/10.1016/j.intell.2018.03.010

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

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31,* 169–180. http://dx.doi.org/10.1177/0146621606291569

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184. http://dx.doi.org/10.1177/01466216970212006

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 1–97. http://dx.doi.org/10.1007/BF03372160

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

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity, 4,* 79–85. http://dx.doi.org/10.1016/j.tsc.2009.06.005

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

StataCorp. (2017). *Stata statistical software: Release 15*. College Station, TX: Author.

Stefanic, N., & Randles, C. (2015). Examining the reliability of scores from the consensual assessment technique in the measurement of individual and small group creativity. *Music Education Research, 17,* 278–295. http://dx.doi.org/10.1080/14613808.2014.909398

Storme, M., Myszkowski, N., Çelik, P., & Lubart, T. (2014). Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges. *Learning and Individual Differences, 32,* 19–25. http://dx.doi.org/10.1016/j.lindif.2014.03.002

Tan, M., Mourgues, C., Hein, S., MacCormick, J., Barbot, B., & Grigorenko, E. (2015). Differences in judgments of creativity: How do academic domain, personality, and self-reported creativity influence novice judges' evaluations of creative productions? *Journal of Intelligence, 3,* 73–90. http://dx.doi.org/10.3390/jintelligence3030073

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577. http://dx.doi.org/10.1007/BF02295596

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492. http://dx.doi.org/10.1177/014662168200600408

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 37–59). New York, NY: Routledge.

Received November 9, 2017
Revision received November 21, 2018
Accepted November 27, 2018 ■