CREATIVITY RESEARCH JOURNAL
https://doi.org/10.1080/10400419.2023.2193040

Routledge
Taylor & Francis Group

ARTICLE

Creativity with 6 Degrees of Freedom: Feasibility Study of Visual Creativity Assessment in Virtual Reality

Baptiste Barbot[a,b], James C. Kaufman[c], and Nils Myszkowski[d]

[a]UCLouvain, Psychological Sciences Research Institute; [b]Yale University; [c]University of Connecticut; [d]Pace University

ABSTRACT

Immersive virtual reality (IVR) takes advantage of exponential growth in our technological abilities to offer an array of new forms of entertainment, learning opportunities, and even psychological interventions and assessments. The field of creativity is a driving force in both large-scale innovations and everyday progress, and imbedding creativity assessment in IVR programs has important practical implications for future research and interventions in this field. Creativity assessment, however, tends to either rely on traditional concepts or newer, yet cumbersome methods. Can creativity be measured within IVR? This study introduces the VIVA, a new IVR-based visual arts creativity assessment paradigm in which user create 3D drawings in response to a prompt. Productions are then rated with modern extensions of a classic product-based approach to creativity assessment. A sample of 67 adults completed the VIVA, further scored using item-response modeling. Results demonstrated the strong psychometric properties of the VIVA assessment, including its structural validity, internal reliability, and criterion validity with relevant criterion measures. Together, this study established a solid proof-of-concept of the feasibility of measuring creativity in IVR. We conclude by discussing directions for future studies and the broader importance and impact of this line of work for the field of creativity and virtual reality.

ARTICLE HISTORY
Received January 31, 2022
Revised February 17,2023
Accepted March 14, 2023

Introduction

Immersive virtual reality (IVR) is rapidly reshaping the psychological assessment and intervention landscape (e.g., Brivio et al., 2020). Capitalizing on its capability to fully immerse people in virtual environments in a "standardized" fashion, IVR has already been used in a range of diagnosis and assessment applications (e.g., Freeman et al., 2017) and numerous treatment and rehabilitation efforts (e.g., Mattila et al., 2020; Tennant, McGillivray, Youssef, McCarthy, & Clark, 2020). These applications include programs that can help treat attention and memory deficits (Faria, Andrade, Soares, & I Badia, 2016) and enhance emotion regulation, well-being, or self-actualization (e.g., Diemer, Alpers, Peperkorn, Shiban, & Mühlberger, 2015). More broadly, IVR-based programs have been viewed as holding great promise for human development in general, in that it has the potential to help people change their viewpoints of the world, and of one another (e.g., Barbot & Kaufman, 2020; Peña & Blackburn, 2013) with lasting effects in the real world (Rosenberg, Baughman, & Bailenson, 2013).

An emerging line of work in which the potential of IVR for human development was recently highlighted is creativity (e.g., Grigorenko, 2019). Creativity is not just about arts and innovation (e.g., Glăveanu, 2014), but, as described herein, it is used in many dimensions of human experience in which we have to solve problems creatively. In a context of greater global challenges, creativity is part of the "human capital" (Walberg, 1988) ever more recognized as a critical asset, from daily problem solving to societal development (Guilford, 1950; Said-Metwaly, Fernández-Castilla, Kyndt, Van den Noortgate, & Barbot, 2020). Indeed, in the midst of particularly difficult times, creativity may hold a key to our ability to survive and thrive (Kapoor & Kaufman, 2020).

After a brief review of the concept of creativity and its benefits for human progress, we highlight the promise of engaging creativity in IVR and the pivotal need for its assessment in such settings. We then introduce a new IVR-based creativity assessment paradigm, and gauge its feasibility through a proof-of-concept in which a classic product-based assessment approach to creativity is applied.

Creativity, its nature and impact on human development

It is easy to assume that creativity is such a nebulous topic that it is impossible to define. Yet, this is

CONTACT Baptiste Barbot baptiste.barbot@uclouvain.be UCLouvain, Psychological Sciences Research Institute, Place Cardinal Mercier 10/L3.05.01, Louvain-la-Neuve 1348, Belgium

a misperception (Cropley, 2015). From the earliest major works (e.g., Guilford, 1950) to currently accepted viewpoints (Hennessey & Amabile, 2009; Plucker, Beghetto, & Dow, 2004), there is solid scholarly agreement. In order for something to be considered creative, it needs to have 2 core components. It should be *original* (i.e., new and different from past work) and *task-appropriate* (i.e., relevant and useful for its designed purpose). If either novelty or effectiveness are missing, then such a product would not be creative (Simonton, 2012). Beyond definitions, there are many theories and models that suggest ways of conceptualizing creativity. Rhodes (1961) introduced the idea of the 4-Ps of creativity: the *Person*, *Process*, *Press*, and *Product*. Much subsequent research has indeed emphasized one of these angles: individual differences associated with creativity, such as cognition or personality (the *Person*), the stages involved in being creative (the *Process*), the impact of environment and context on creativity (the *Press*), and how to evaluate or determine what is creative (the *Product*), which is one of the key focus of the work presented here.

Whether focused on the *Person*, *Process*, *Press*, or *Product*, creativity is surely a promising target for a new era of digital applications given its multiple individual and social benefits. Indeed, creativity is, has been, and will continue to be a key driver of human progress; most of the major, paradigm-shifting innovations that have enriched our lives are the result of creative genius (Simonton, 2009). Further, much of our economic progress and organizational successes are also rooted in exceptional creative accomplishments (Florida, 2014). However, everyday creators (i.e., everyone) can also receive a myriad of emotional, mental, and physical benefits of creativity. People who are more creative, for example, see a host of possible advantages that range from being better equipped to grow after living through a traumatic experience (Forgeard, 2013) to being seen as more sexually attractive (Kaufman et al., 2016). Engaging in creative activities (typically in the arts) has a wide array of ways in which it improves people's lives. Some are existential, such as making sense of their past life events (Pennebaker & Seagal, 1999) or finding meaning (Kaufman, 2018). Other ways are of cognitive nature. For example, older adults who participate in creative arts are less likely to develop dementia (Roberts et al., 2015) or to stave off its effects if they already have dementia (Maguire, Wanschura, Battaglia, Howell, & Flinn, 2015). Creativity has also established impact on improved mental health outcomes across many dimensions, from reducing stress (Meier et al., 2020), to relieving one's burdens (Goncalo, Vincent, & Krause, 2015). Finally, creativity has potential to encourage helping others and reduce prejudice (Groyecka-Bernard, Karwowski, & Sorokowski, 2021; Luria & Kaufman, 2017), as well as supporting the development of self and identity (Barbot & Heuser, 2017; Barbot, 2020, 2021; Sica, Ragozini, Palma, & Sestito, 2019).

Assessing creativity: bringing creativity in digital environments

Digital worlds offer countless new modalities for creative expression (Barbot, 2021; Hoffmann, Ivcevic, & Brackett, 2016) and there is an emerging line of work aiming at stimulating creativity through digital technologies (Tang, Mao, Naumann, & Xing, 2022), including in IVR (Chang, Kao, & Wang, 2022; Chen, Chang, & Chuang, 2022; Graessler & Taplick, 2019; Guan, Wang, Chen, Jin, & Hwang, 2021; Lau & Lee, 2015; Li et al., 2022; Nelson & Guegan, 2019; Obeid & Demirkan, 2020; Ritter et al., 2012; Thornhill-Miller & Dupont, 2016; Wang, Weng, Tsai, Kao, & Chang, in press; Yang et al., 2018).

This body of work has explored the use of IVR as a tool to enhance creativity in problem-solving (Yang et al., 2018; for review, see Graessler & Taplick, 2019), education (Guan et al., 2021), engineering (Graessler & Taplick, 2019), or design (Chang et al., 2022; Obeid & Demirkan, 2020). Some of this work examined the effect on creativity of either unusual IVR environmental (Lau & Lee, 2015) or simulated "real" environments that are conductive of creative thinking (Li et al., 2022; Nelson & Guegan, 2019). All these studies concluded on the promise of IVR for promoting creativity in various settings (e.g., Li et al., 2022; Thornhill-Miller & Dupont, 2016). They also pointed out how VR can increase motivation and engagement in the task (e.g., Guan et al., 2021; Lau & Lee, 2015). However, they have all pointed out the need for further research to fully understand the effects of IVR on creativity.

Surprisingly, there are, however, scarce attempts to measure creativity in such settings: most research in the field of creativity in IVR typically measures aspects of the creative potential outside the virtual environment using classic methods (e.g., Chen et al., 2022; Wang et al., n.d.). Conducting creativity assessment outside the IVR settings may pose experimental and practical challenges (e.g., constraints on the timing of the assessment; having to set up a participant for IVR multiple times to conduct activities both inside and outside IVR). After some early experimentations (Ward & Sonneborn, 2009), there has indeed been strikingly little scholarship on creativity measurement within virtual reality settings. Given the paucity of funding for creativity research (Runco & Abdullah, 2014), this omission is in fact not surprising; there has been scant work on applying modern technologies such as computer scoring to actual creative performance, despite recent efforts (e.g., Johnson et al., 2022; Zedelius, Mills, & Schooler, 2019). The lack of creativity assessment research in the context of IVR is particularly detrimental for multiple reasons. First, as outlined since decades, VR environments offer optimal context for assessment, providing more controlled and ecologically valid measurement settings (e.g., Blascovich et al., 2002). Second, when the specific assessment purpose is to measure creativity, IVR seems particularly relevant to let users creatively deal with unusual stimuli, situations, and modalities that they have never been exposed to before. Third, IVR environments seem particularly relevant to gauge or monitor change in creativity in the context of IVR-based creativity stimulation programs that are gaining momentum (e.g., Barbot & Kaufman, 2020; Nelson & Guegan, 2019; Ritter et al., 2012). Together, imbedding creativity assessment within IVR (in the context of IVR-based studies, interventions, or training programs) offers the practical advantage to incorporate the assessment component within the same environment at any time in the experimental setup.

In the work presented here, we propose to transpose classic assessment of creativity in the realm of IVR. In fact, many current creativity assessments are still based on core concepts introduced in the 1950s (see e.g., Barbot, Hass, & Reiter-Palmon, 2019; Kaufman, Baer, Cole, & Sexton, 2008). One very established method in the field – and comparatively more recent – is the Consensual Assessment Technique (CAT; Amabile, 1996). Although there is a range of variations as to how it is applied in creativity research (Cseh & Jeffries, 2019), its basic principle remains the same. It consists of collecting a corpus of work from the participants to be tested. Such products can be nearly anything, from artwork to business ideas, to mathematical equations, musical compositions or written productions (Barbot & Lubart, 2012). A group of judges generally selected for their expertise in that specific domain is then involved. Using their own conceptions of creativity, they independently evaluate all products (Kaufman & Baer, 2012) or some sub-samples of products using planned missingness designs (Barbot, 2020; Fürst, 2020). After statistically gauging the consensus across judges, scores are then aggregated – generally using a mean or sum score which is not without limitations (Myszkowski & Storme, 2019a) – so that each production may be characterized by a composite creativity score.

In CAT studies, judges with expertise in the relevant area consistently yield high inter-rater agreement (Baer, Kaufman, & Gentile, 2004; Baer, Kaufman, & Riggs, 2009). In contrast, it takes many more novices to reach a basic level of reliability, and their judgments do not particularly correlate with those of experts (Kaufman, Baer, & Cole, 2009; Kaufman et al., 2008). For domains with less established guidelines or required knowledge – such would be the case for most tasks developed for IVR – advanced students (Kaufman, Baer, Cropley, Reiter-Palmon, & Sinnett, 2013) and dedicated aficionados of the domain (Plucker, Kaufman, Temple, & Qian, 2009) show notably better reliability and agreement with expert opinion.

The CAT is frequently used in creativity research (Forgeard & Kaufman, 2016) across a range of domains. For instance, it is commonly used to evaluate creative writing (Taylor et al., 2021), but examples in the visual art are scarce and even more so in unusual settings such as with 3D-Drawing applications in IVR environments. The latter is of particular interest because many people might never have had experience creating in such settings. Therefore, it also represents a unique opportunity to elicit people's "baseline" creative behaviors as they discover a new, unusual creative situation.

Present study

Given such gaps, it seems to be time to move creativity assessment into the 21st century and incorporate the promise of IVR in this pursuit. In an effort to move this line of work forward, the present study represents a first attempt to transpose classic creativity measurement methods to IVR settings. Specifically, we sought to address whether digital creative production completed in IVR settings by people unexperienced with the expressive modality at hand (here, 3D painting) would be suitable for assessment under a classic creativity assessment paradigm, namely, the Consensual Assessment Technique (CAT; Amabile, 1996). Would such assessment setting with realistic testing constraints (i.e., prompt, time limit) in an environment largely unknown to the user, be measured with a sufficient level of structural validity (i.e., supporting an objectifiable reality), and yielding creativity scores internally reliable? Further, would individual differences in "classic" expressive creativity task (e.g., writing), relate to creativity measured in this new, digital modality? Finally, would the creative quality of the digital productions generated in this IVR modality be sufficiently distinguishable from other aspects of the productions, including their aesthetic qualities? As an initial proof-of-concept, the present study addresses these questions with a rather simple paradigm involving creative production in first-time 3D-painting users, further rated by applying the classic CAT methodology.

Method

Participants and procedure

The present study was part of a research program on the development of creativity and imagination in IVR, approved by Pace university IRB protocol 16-130. Participants were 67 undergraduate students ($n_{male} = 35, n_{female} = 31, n_{other} = 1, M_{age} = 20.97,$ $SD_{age} = 4.82$) recruited from a large urban University in the Northeastern region of the United States. They represented a diverse range of racial and ethnic background including Caucasian/White (51%), Asian American (21%), Multiracial (15%), African American/Black (12%) and Pacific Islander/Native Hawaiian (1%). Twenty-one percent of the participants indicated being of Hispanic background. Most of them (77%) reported having some college attendance (but no college degrees), and a variety of disciplines were represented. As described in greater length in related work, participants were recruited through in-class announcements, flyers, and IVR showcase events conducted in the main lobby of the university. Besides a brief preview of HMD technology[1] that those participants recruited through the IVR event had, none of the participants have had IVR experience prior to their involvement in the study.

After providing consent to participate, eligible subjects were enrolled in a 5-weekly sessions Immersive Virtual Reality Experience (IVRE) program (Barbot & Kaufman 2020). One week prior to the start of the program, they completed a series of measures online including the *Visual Aesthetic Sensitivity Test – Revised* (VAST – R) (cf. 2.3. measures), and as part of the first IVRE session, they completed another set of in-lab measures (including the *Storyboard Task*), as well as the Virtual Immersive Visual Art (VIVA) Task under focus here (2.2. Material). After the last IVR session, they completed another set of measures (including an alternate form of the *Storyboard Task*).[2] At the end of the program, participants received a $25 gift card for an online store as a token of appreciation for their time.

Materials

The VIVA Task was implemented in the award-winning application *Tilt Brush* (Google, 2016). It was delivered individually to participants in a comfortable 3 m (9 ft 10.1in) × 3 m (9 ft 10.1 in) room-scale settings, using the HTC Vive HMD, which provides a 110-degree field of view at 90 Hz refresh rate. The audiovisual rendering was provided to the wired HMD using a PC with a 3.00 GHz AMD 8 × 3.00 GHz processor, 16 Gb of RAM, and a 4 Gb VRAM Nvidia GeForce GTX 980 graphics card. Tilt Brush is a popular room-scale 3D-painting IVR application with a 6 degrees of freedom (6DoF) motion interface, allowing both rotational and translational user's movements in IVR. One of 2 HTC Vive handheld controllers is given to the participants to access a virtual palette from which they were able to select from a broad range of painting tools. The virtual brush materialized by the other handheld controller is tracked in real-time. When using the controller's trigger, participants' movements of the controller create brush strokes that follow in the 3-dimensional canvas (Patz, Grzymala, & Gengnagel, 2019).

Measures

VIVA task: creativity and aesthetic quality

Once the *Tilt brush* application was launched and basic information on controls was provided, participants were prompted to "*draw an animal of your choice, real or imaginary.*" Participants were given 12 minutes to complete their production in the IVR environment. Once participants were finished, their production was saved and screenshots taken for follow-up evaluation. Specifically, productions generated from the VIVA task were rated for both creativity and aesthetic quality by 9 judges with a range of creativity expertise, using the above described Consensual Assessment Technique (CAT; Amabile, 1996), as implemented in CAT-i (Barbot, Orriols, & Pouyade, 2008)—a web-based interface to facilitate CAT procedure online. Ratings were provided independently by each judge for each production presented in a random order, using 7-point Likert-type scale, with "1" denoting low creativity, and "7" denoting high creativity. Although the central purpose of CAT is to measure the creativity of productions, it was also used elsewhere (e.g., Amabile, 1996) and here to collect ratings on the aesthetic quality of the productions, in order to ensure that the measure of creativity being proposed was empirically distinguishable from a measure of aesthetic abilities (i.e., discriminant validity). Ratings were obtained in 2 separate rating sets, 1 for creativity, and 1 for aesthetic quality.[3] The 'Data analyses and statistical approaches' section extensively discusses how the ratings were analyzed and production scored. The same procedure was applied separately to both the creativity and the aesthetic quality sets of ratings. Figure 1 illustrates sample productions that were generated by the participants. Examples were selected to represent a range of creativity and aesthetic quality levels, based on ratings of these productions (see the 'Data analyses and and statistical approaches' section).

Creative writing task

The Storyboard task (Taylor, Kaufman, & Barbot, 2021) was used as another production-based assessment of creativity, with a focus on the writing domain with a narrative framework, that is, "essentially the introduction of a problem, with particular characters and setting, which includes a beginning, a middle and an end" (Barbot, Tan, Randi, Santa-Donato, & Grigorenko, 2012). The choice of a writing task, as opposed to a graphic task, as an external validity criterion was intended to provide insights on the domain-general aspects of creativity elicited by VIVA. Although both the VIVA and the storyboard task can be categorized as "expressive" creative tasks, they theoretically rely on very distinct domain-specific skills. As such, their common variance should reflect more domain-general aspects of creativity. The task includes a total of 4 items across 2 forms, delivered on a digital platform, each of which presents a set of 3 unrelated images, prompting participants to *"Write a story using each picture as an illustration of the beginning, middle, and end of your story."* Accordingly, for each item, participants type out each part of their story in 3 separate text boxes underneath the corresponding picture, in a self-paced format. Following classic CAT procedure, 3 independent expert judges rated the products of both tasks on their level of creativity. A few participants (between 2 and 5 depending on the measurement occasion) did complete all storyboard items, but all participants had scores on at least 2 items.

Because the measurement design involved multiple raters and items, we used a Generalized Many Facet Rasch Model (GMFRM), estimated with the R package "sirt" (Robitzsch & Steinfeld, 2018) to obtain factor scores for this task. Many Facet Rasch Models (MFRM) are item-response models that are designed to accommodate for simultaneously the person (participant), the rater, and the item as sources of variability in ordinal ratings. They are extensively discussed elsewhere (see Primi, Silvia, Jauk, & Benedek, 2019 for a discussion of their use on CAT data) and have been used in similar contexts, including with smaller sample sizes as in the present study (e.g., Barbot et al., 2012; Tan et al., 2015). Contrary to previous uses in this context, we here used a Generalized variant (see Robitzsch & Steinfeld, 2018), which had the advantages of (a) not assuming all raters and items to be equally discriminant – making it more realistic than the classic MFRM, since the raters were of varying expertise level – and (b) better fitting the data than the MFRM – $\Delta\chi^2(3) = 33.29, p < .001$. The estimated reliability based on the model was satisfactory at .92. For comparison purposes, the Cronbach' α across items also yielded

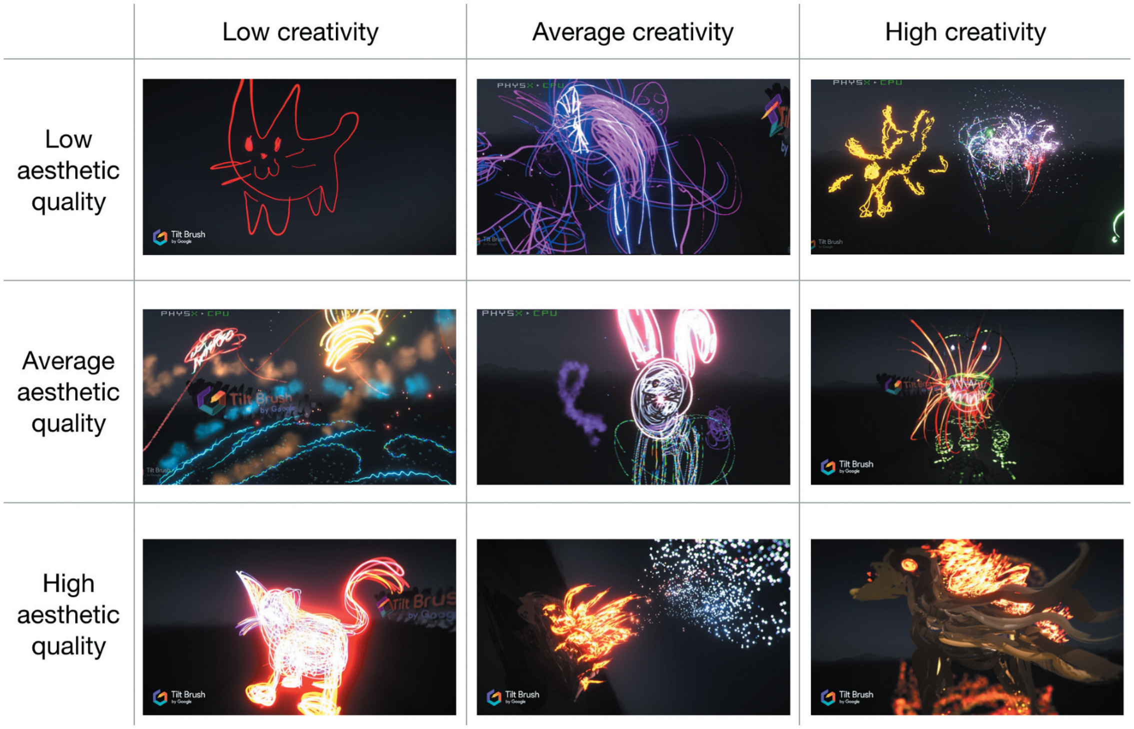

Figure 1. *Sample productions on the VIVA task.* VIVA – Virtual Immersive Visual Art.

excellent inter-rater agreement (Cronbach's $\alpha = .96$, .94, .90, and .90 respectively for items A1, A2, B1, and B2), as well as the average scores of the 4 items (Cronbach's $\alpha = .86$).

Visual aesthetic sensitivity

Aesthetic abilities were measured in order to further support the VIVA task's discriminant validity. To do so, the Visual Aesthetic Sensitivity Test – Revised (VAST – R; Myszkowski & Storme, 2017) was employed. The VAST-R measure one's ability to make aesthetic judgments in agreement with external standards – generally defined as aesthetic sensitivity (Child, 1964; Myszkowski, 2020). It consists of 25 items, which are a subset of the 50 original items (VAST; Götz, 1985), each consisting of a pair of black and white abstract drawings by the German painter K. O. Götz, Each pair (i.e., item) comprises an original drawing, and a version of that drawing that is altered to include aesthetic defects (e.g., line breaks). The participants are prompted to indicate which of the 2 drawings in the pair is aesthetically superior or better balanced. The items or the original VAST (and thus VAST – R) were selected based on unanimous agreement by a panel of 8 art experts over the "correct" responses. Although aesthetic quality is certainly subjective, the agreement of experts forms an empirical standard, which is a practical necessity for this type of test (Myszkowski, Çelik, & Storme, 2020). In the present study, the observed Cronbach's α (or KR20), based on the tetrachoric correlation matrix (the items being dichotomous), was .81, indicating an internal reliability that was both sufficient and comparable with previous research (Myszkowski & Storme, 2017; Myszkowski, 2019a).

Data analyses and statistical approaches

Analyses were conducted in 3 main sets focused on examining the psychometric properties of the VIVA Task scores including (a) their structural validity, (b) internal reliability, and (c) criterion validity with the criterion measures.

Structural validity

Given that the CAT judgment data are ordinal, we followed recent recommendations (Myszkowski & Storme, 2019b; Myszkowski, 2019b) and emphasized a psychometric modeling approach based on ordinal item-response models. This approach is comprehensive in that it allows to compare the fit of various psychometric models and estimate the resulting reliability of scores provided by each model. We used the R package "jrt" (Myszkowski, 2019b) for these analyses, which itself is based on the estimation engine of the "mirt" package (Chalmers, 2012; Liu & Chalmers, 2018). "jrt" tests 8 response models extensively described in Myszkowski (2019b) and the item-response theory literature (e.g., Nering & Ostini, 2010), including the popular Graded Response Model (GRM; Samejima, 1969), the Generalized Partial Credit Model (GPCM; Muraki, 1990), the Partial Credit Model (PCM; Masters, 1982), and the (Rasch) Rating Scale Model (RSM; Andrich, 1978). The most flexible models (the GRM and GPCM) make minimal assumptions about the judgments, as they allow for judges to differ in difficulty/severity, discrimination/expertise (ability to capture in their judgment the latent attribute measured), and use of the response scale (structure of the responses categories 1-2-3-4-5-6-7).

In the present study, "jrt" determined that 4 models could not be fitted to the data because all raters did not use all response categories. As a result, only the 4 remaining models we tested (i.e., GPCM, PCM, GRM, and a GRM variant with all judges having equal discriminations, thereafter labeled Constrained Graded Response Model [CGRM]). In addition to fitting these item response models, the R package "lavaan" (Rosseel, 2012) was used to fit a unidimensional Covariance-based Confirmatory Factor Analysis (CFA) with Maximum Likelihood estimation, which is a popular modeling approach of CAT data (e.g., Barbot, 2020).[4] Compared to the ordinal item-response models, this procedure has the disadvantage of assuming multivariate normality, but the advantage of relying on less judge parameters, thus allowing a more stable estimation.

For all models, classic structural model fit measures were computed (Cai & Hansen, 2013) and their established benchmark applied (Hu & Bentler, 1999) including a non-significant χ^2 tests, the Comparative Fit Index (CFI) and the Tucker–Lewis Index (TLI) with values above .95, the Root Mean Squared Error of Approximation (RMSEA) and Standardized Root Mean Squared Residual (SRMR) below .08, to suggest a satisfactory model fit. Nested models (GRM/CGRM and GPCM/PCM) were compared with Likelihood Ratio (χ^2) Tests, and non-nested models were compared using the Akaike Information Criterion (AIC), with smaller AIC indicating better fit.

Internal reliability

In item-response models, reliability is a function of the latent attribute. For example, extreme creativity levels may be less reliably measured than average-range creativity levels. Accordingly, we computed and plotted the reliability function, as implemented in “jrt.” We also computed the empirical reliability, which is the average reliability in the sample, as well as the expected reliability based on a prior standard normal distribution of creativity levels. In addition to these analyses based on ordinal item-response models and for comparison purposes, we also computed the Cronbach α which is the most common way to evaluate CAT scores’ internal reliability (Cseh & Jeffries, 2019) despite severe limitations (Myszkowski & Storme, 2019b).

Criterion validity

To study criterion validity, we computed bivariate correlations between the factor scores estimated for all productions – from the best fitting model identified in the structural validity analytic set – and the criterion measures scores (i.e., the Storyboard creative writing, and VAST-R scores). Consistent with extent literature, and psychometric evaluations data of other product-base creativity assessment (e.g., Barbot & Lubart, 2012; Kaufman, Lee, Baer, & Lee, 2007; Taylor & Barbot, 2021), we expected that the VIVA Creativity scores would be positively correlated with creativity levels in the Storyboard task, aesthetic quality ratings of the VIVA productions, and VAST-R scores. Further, we expected that these associations would be of moderate magnitude at most ($r < .5$), indicating that the VIVA Creativity task would still be empirically distinguishable from these other measures.

Results

Structural validity

The fit indices of the various models tested are reported in Table 1. They indicate that overall, all ordinal models fit the data very well. However, the least constrained models (the GRM and GPCM), which do not assume all judges to be equally discriminant, had a better fit than their constrained counterparts. Specifically, for creativity ratings, the GRM fit significantly better than the CGRM – $\Delta\chi^2(8) = 24.37, p = .002$– and the GPCM fit significantly better than the PCM – $\Delta\chi^2(8) = 17.51, p = .025$. The GRM and GPCM are not nested models and thus could not be compared with a Likelihood Ratio Test, but the lower AIC associated with GPCM suggested that the latter had better fit. Similar results were found for the aesthetic quality ratings, with the GRM fitting significantly better than the CGRM – $\Delta\chi^2(8) = 24.20, p = .002$ – the GPCM fitting significantly better than the PCM – $\Delta\chi^2(8) = 18.62, p = .017$ – and the GPCM having a slightly better fit than the GRM based on their AICs. Finally, for the sake of comparison, the traditional (covariance-based) CFA model returned an excellent fit to the data for creativity – $\chi^2(27) = 28.35$, p = .393, CFI = .994, TLI = .992, RMSEA = .027, SRMR = .056) – and acceptable fit for aesthetic quality – $\chi^2(27) = 40.53$, p = .046, CFI = .941, TLI = .922, RMSEA = .087, SRMR = .069). Overall, the ratings being ordinal in nature, we chose the best fitting ordinal model, which was the GPCM. Although this model is extensively described elsewhere (e.g., Muraki & Muraki, 2016), it is notable here that it allows for judges to vary in their level of severity, their discrimination (i.e., expertise) and the structure of the response categories, thereby also making it a realistic model, since the raters received no training and thus had no reason to be equally severe or discriminant, nor to use the response scale in the same way.

The best fitting models (the GPCM for both creativity and aesthetic quality) were used to derive the Judge Category Curves, which plot the predicted probabilities that a judge chooses a response category on the 1–7 rating scale, as a function of the latent attribute (e.g., creativity). These plots are presented in Figure 2 for creativity ratings and in Figure 3 for aesthetic quality ratings. The x-axis represents θ, the latent attribute being judged (i.e., creativity or aesthetic quality), which is represented on a standard score (“z”) scale (i.e., with mean = 0 and SD = 1). The y-axis represents the predicted probability that the category is

Table 1. Fit Indices of the Different Ordinal Response Models

Ratings	Model	χ^2	*df*	*p*	CFI	TLI	SRMR	RMSEA	AIC
Creativity	CGRM	42.53	35	.178	.982	.982	.110	.057	1968.9
	GRM	24.34	27	.612	1.00	1.00	.072	.000	1960.6
	PCM	46.06	35	.100	.974	.973	.099	.069	1961.4
	GPCM	26.90	27	.469	1.00	1.00	.064	.000	1959.9
Aesthetic quality	CGRM	57.24	35	.010	949	.948	.130	.099	1930.1
	GRM	32.58	27	.211	.987	.983	.092	.056	1921.9
	PCM	56.90	35	.011	.950	.949	.120	.098	1920.4
	GPCM	32.35	27	.219	.988	.984	.087	.055	1917.8

CFI – Comparative Fit Index; TLI – Tucker–Lewis Index (values above 1 truncated to 1); SRMR – Standardized Root Mean Square Residual; RMSEA – Root Mean Square Error of Approximation; AIC – Akaike Information Criterion.

chosen and the different lines represent the different response categories (i.e., 1 through 7). Both figures clearly indicate variability in severity and use of the response scales across raters. For example, Judge 4 overwhelmingly used the response category 2, judge 7 underused the response category 4, and Judge 5 was very severe, most likely giving "1" for any production of lower than average creativity. Such variability supports the use of ordinal models such as the GPCM used here to estimate creativity and aesthetic quality, as opposed to using sum/average scores, which, unrealistically assume all judges to be equally discriminant and to use the response scale similarly.

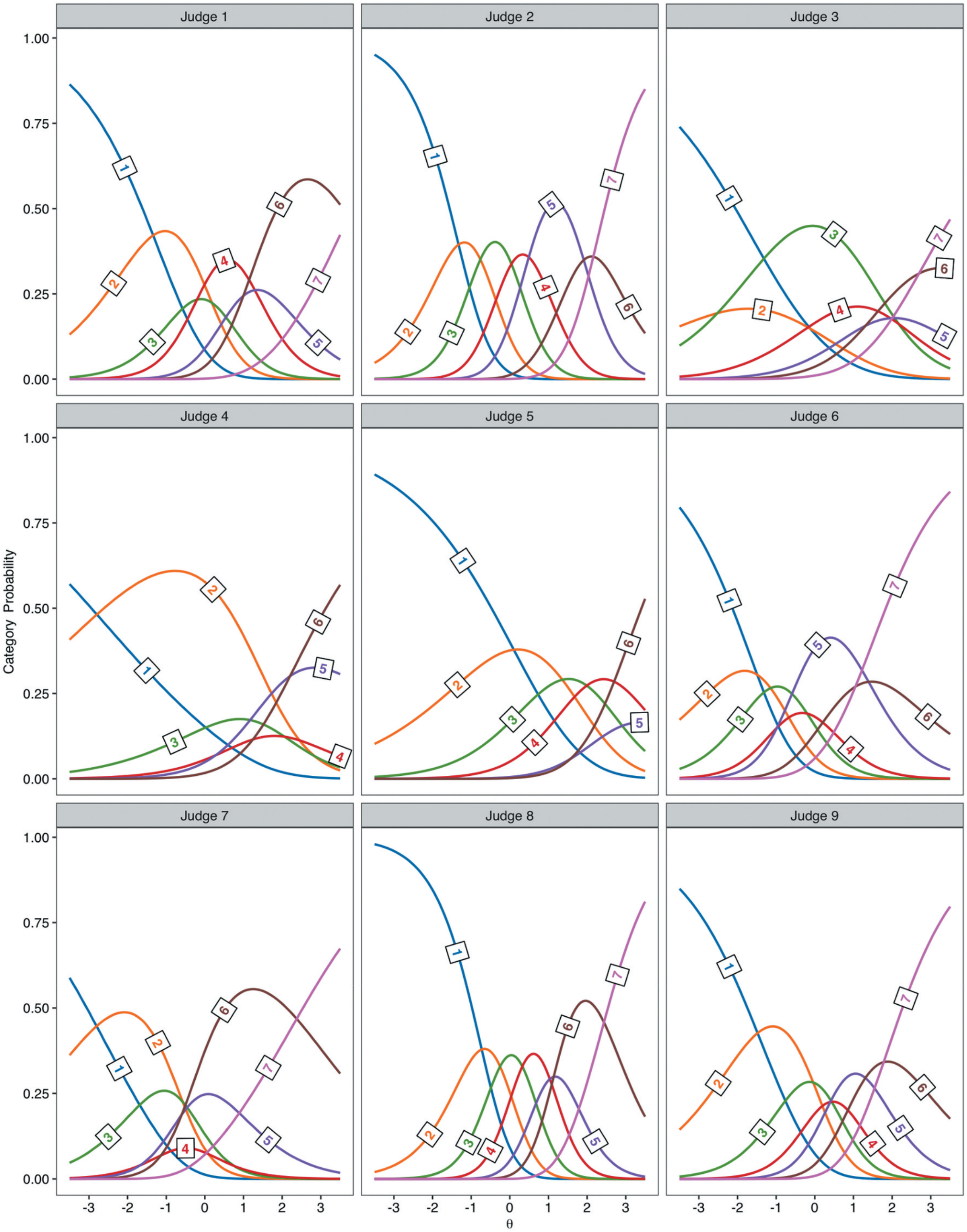

Figure 2. Judge category curves for the creativity judgments.

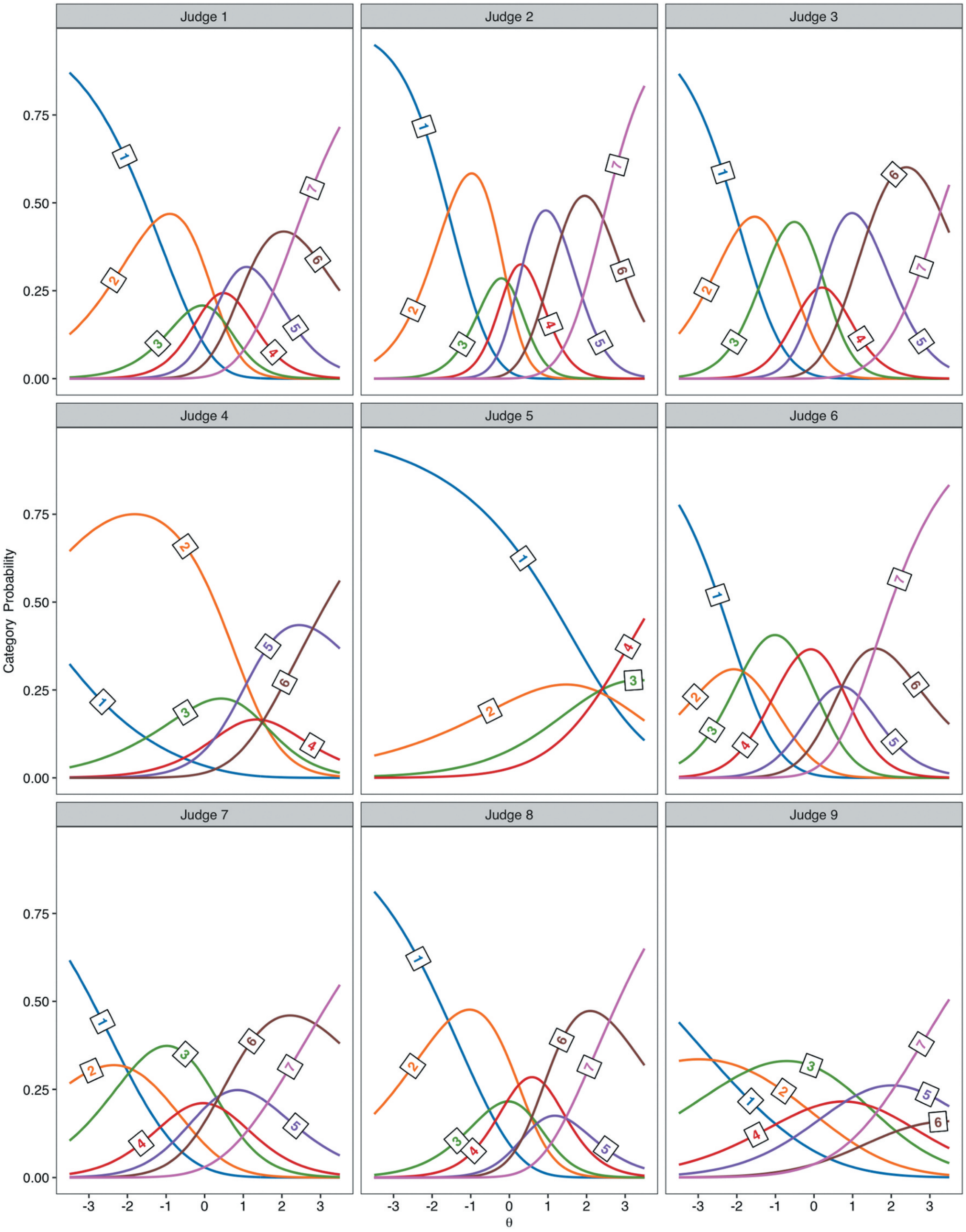

Figure 3. Judge category curves for the aesthetic quality judgments.

Internal reliability

In item-response models, reliability is conceptualized as conditional upon the levels of latent attribute. We therefore present reliability functions for the creativity and the aesthetic quality ratings respectively in Figure 4 panels A and B, based on the best fitting models (the GPCM in both cases). Like previously, the x-axis represents the latent attribute being rated (i.e., creativity and aesthetic quality, respectively). However, the y-axis here represents the total reliability of the set of ratings. The inspection of these reliability functions indicated that

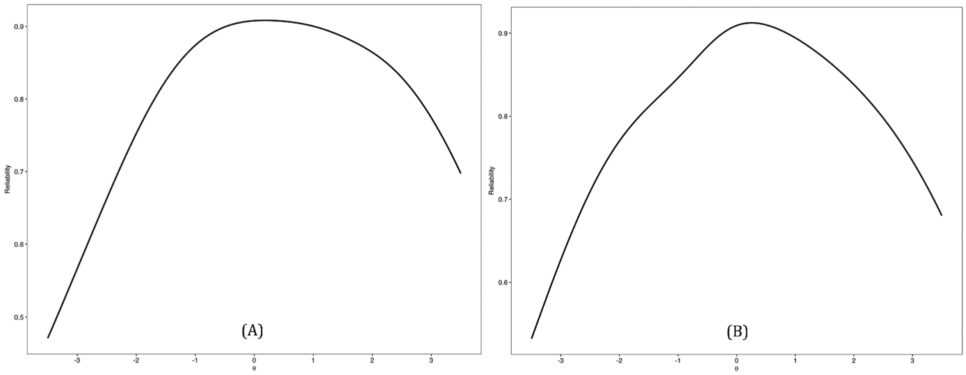

Figure 4. *Conditional reliability of the creativity* (panel A) *and aesthetic quality ratings* (panel B).

satisfactory reliability was obtained for creativity and aesthetic quality levels that were around average and that better reliability was achieved for high levels of creativity and aesthetic ability compared to low levels.

The average reliability in the sample – often named empirical reliability – was .89 for creativity ratings and .86 for aesthetic quality ratings. The expected reliability based on prior standard normal distributions of creativity and aesthetic quality was similar, with .89 for the creativity ratings and .90 for the aesthetic quality ratings. Cronbach's α, which was computed for the sake of comparability with other studies, was .87 for both the creativity ratings and the aesthetic quality ratings. These results are very consistent, and clearly indicate satisfactory internal reliability. Because of the satisfactory fit of the GPCM and the reliability of the ratings, we computed factor scores based on the GPCM model, using Expected A-Priori (EAP) estimates, for both creativity and aesthetic quality for use in subsequent analyses.

Criterion validity

Bivariate correlations between the study measures and age (for control purpose) are presented in a correlation matrix (Table 2). Overall, as hypothesized, the pattern of correlations observed suggested that the VIVA Creativity scores were correlated with, but distinguishable from, the storyboard creativity task and the VAST – R. Specifically, VIVA creativity was moderately related to written creativity (i.e., Storyboard task; $r(65) = .45, p < .001$) and independent of aesthetic abilities (i.e., VAST-R; $r(65) = .21, p = .096$). However, VIVA creativity and aesthetic quality were strongly related to each other ($r(67) = .85, p < .001$), whith the latter showing a smaller association with both written creativity (i.e., Storyboard task; $r(65) = .37, p = .002$) and aesthetic abilities (i.e., VAST-R; $r(65) = .26, p = .033$).

Discussion

The present study represents a first attempt to bridge a gap between an enthusiastic line of IVR-based programs and the ever more crucial need for creativity, a fundamental aspect of human capital. As a proof-of-concept, we proposed a rather simple assessment paradigm in IVR which builds upon the unique capabilities of IVR technology and modern approaches to the classic CAT method. This paradigm consists of asking users to create a visual art digital production in IVR, using a 3D-painting application – a production setting never experienced before by most users – to capture their "baseline" approach to creatively expressing themselves in unusual settings. This study has clearly demonstrated that this proposed setting as a context for creativity assessment is feasible. All participants were – to varying extent and with varying level of "fluency" – able to

Table 2. Correlations Between the VIVA Task and External Criteria

	M	SD	1	2	3	4
1. Age	20.97	4.82				
2. VIVA Creativity (factor scores)	0.00	0.95	–.04			
3. VIVA Aesthetic Quality (factor scores)	–0.01	0.94	.02	.85***		
4. Storyboard Creativity Task (factor scores)	–0.29	1.29	–.03	.45***	.37**	
5. VAST – R (item success rate)	0.77	0.13	.02	.21	.26*	.09

*$p < .05$, **$p < .01$, ***$p < .001$.

generate such productions in IVR settings. After discussing specific findings regarding the psychometric evaluation of this new assessment paradigm as well as some study limitations, we conclude by discussing directions for future studies and the broader importance and impact of this line of work for the field of creativity and virtual reality.

Psychometric evaluation

Structural validity and internal reliability

The measurement procedure proposed here appeared to yield satisfactory psychometric properties, in that the adequate fit of the tested unidimensional response models supported that the measurement of creativity was indeed unidimensional. In other words, when asking judges to rate for creativity, they appear to have all identified the same quality (Amabile, 1996). Correspondingly, model-based reliability estimates – as well as Cronbach's α classically employed in CAT – indicated satisfactory reliability of creativity scores. Of note, the conditional reliability estimates for creativity (Figure 3) suggested that most productions could be rated with very high precision across raters; only those productions with extremely low levels of creativity were assessed with lower reliability. This finding may be due to the raters not being able to discriminate among low and lower creativity/aesthetic quality levels. We could speculate that this lower reliability at low levels could be related to some VIVA productions that may have been simply harder to understand, and thus harder to judge – a bit like judging the creativity and aesthetic quality of poems that are not comprehensible or are in a foreign language. The same goes for the aesthetic quality dimension captured here, revealing the same pattern of internal reliability across aesthetic quality levels, as well as support for its (unidimensional) structural validity. While the focus of the present work was mainly to combine the VIVA Task with the Consensual Assessment Technique as a new, IVR-based creativity assessment paradigm, the satisfactory psychometric properties observed across both creativity and aesthetic quality ratings suggest that the VIVA procedure may also be recommended for use as a measure of other IVR-based drawing skills (and possibly, other relevant criteria).

On related note, a brief comparison of Figures 2 and 3 shows great similarities in the way judges rated the 2 attributes (e.g., per previous illustration, judge#4 use of the response "2," judge#5's severity, are all consistent across attributes) suggesting that raters characteristics (severity, use of the response scale, discrimination) may be relatively stable across the drawing attributes being assessed. This observation of rating style consistency has not been heavily investigated as such in the CAT literature and deserves further investigation, considering both practical – e.g., could help facilitate model estimation by applying common parameters across multiple rating sets/criteria – and substantive implications (i.e., generalizability of judges rating style).

For both VIVA drawings and storyboards, response models that allowed for raters to vary in discrimination fit the data significantly better than models that did not. This indicates that the raters significantly differed in discrimination (i.e., expertise) in the ratings regardless of the task (VIVA or storyboard) and regardless of the attribute judged in the VIVA task (creativity or aesthetic quality). Although the sample size is rather small and although this may be explained by a lack of training of the raters, this result suggests that one should not assume raters of creativity tasks to be equally discriminant (or, in the Classical Test Theory sense, tau-equivalent), when analyzing rating data, because this assumption is likely unrealistic. This is in line with recent research (Myszkowski & Storme, 2019b) that questions the use of Cronbach's α and sum scoring – which assume tau-equivalence – in similar contexts. In item-response modeling, it also suggests to consider methods that generalize Rasch modeling by allowing for variable rater discrimination (i.e., using models like the Generalized Partial Credit Model, Graded Response Model or, in multiple item and rater contexts, a Generalized Many Facets Rasch Model).

Criterion validity

With respect to criterion validity, the pattern of correlations observed between VIVA creativity scores and other tasks were consistent with theoretical expectations, indicating satisfactory criterion validity. Indeed, given the partly domain-specific nature of creativity, it was not expected to observe very strong correlations between creativity scores belonging to productions in different domains, or even, between distinct tasks across domains (Baer & Kaufman, 2005; Barbot, Besançon, & Lubart, 2016). However, Barbot (2020) reported a (latent, i.e., deattenuated) correlation of .44 between a drawing composition task, and a story writing task. Here, we too observed a moderate, but larger than usual (e.g., Avitia & Kaufman, 2014) association between creativity measured with the storyboard task and the VIVA task ($r = .45, p < .001$) supporting a good convergent validity between these tasks, and consistent with Barbot (2020)'s findings using 2 paper-and-pencil production tasks, in drawing and writing, respectively. The magnitude of this correlation is surely attributed to some domain-general components of creativity, including effort and persistence – for which the story board task is very sensitive (Taylor & Barbot, 2021) – and which might further explain, to some extent, the unexpected association between storyboard creativity and VIVA aesthetic quality ($r = .37, p < .001$).

Relatedly, the correlation between Aesthetic Quality and Creativity in the VIVA Task was stronger than expected, suggesting that the creativity ratings collected in the task largely captured the aesthetic quality of the drawings. There are several sets of explanation for this finding. First, this large association is likely inflated by a common method bias, for the reason that the same judges have rated creativity and aesthetic quality. In CAT literature, it is not unusual to involve different raters for different qualities of the production, avoiding this common method bias inflation (Amabile, 1996). Second, it is possible that limited expertise among the raters for the task at hand might have contributed to confounded both criteria (Kaufman & Baer, 2012). Perhaps visual-art experts (either artists or teachers) would have been more discriminant in separating both criteria.

Helping to disentangle this pattern of association, it is to be noted that VIVA creativity was not significantly correlated with the measure of aesthetic ability (VAST – R), supporting the fact that both measures tap into distinct constructs. In contrast, the VIVA aesthetic quality was significantly and positively correlated with the VAST – R. Although secondary to the present investigation, this finding is consistent with previous empirical (Myszkowski, Storme, Zenasni, & Lubart, 2014) and theoretical accounts (Myszkowski & Zenasni, 2016; Myszkowski, Storme, & Zenasni, 2016), which suggest that visual aesthetic sensitivity is a useful predictor of creative abilities in the visual domain. This is also consistent with findings (e.g., Kozbelt, Seidel, ElBassiouny, Mark, & Owen, 2010) indicating that visual artists tend to have higher visual abilities. In all, what is particularly striking with the present findings is that the IVR modality was new for all the participants, and 3D-painting experience surely different than any other visual art experiences they might have had in other contexts. Given that the resulting VIVA products were judged reliably and displayed evidence of validity, this suggests that people reasonably transfer their creative and aesthetic skills in the IVR modality.

Limitations

Overall, the present study represents a promising proof-of-concept and suggests meaningful directions, but it should be considered in light of several limitations. First, for the sake of simplification, we have offered a single VIVA task item, namely, the creation of an animal. Many other tasks of this nature could be derived, and it would be useful to gauge consistency of performance across items or alternate forms – a recurrent issue in creativity assessment, particularly for the study of creativity change and development (e.g, Barbot, 2019). Present evidences of consistency across 2 domains suggest good hopes on that matter, but it will need to be established more formally. Second, although creativity and aesthetic quality might represent the "building blocks" of visual art creativity (Pelowski, Leder, & Tinio, 2017), many more aspects of creative production in IVR environments could have been accounted for. For instance, we could have considered many special features of the production (e.g., use of space, extent of 3-dimensionality of the production) taping perhaps on IVR-specific skills not necessarily involved in other visual-art tasks. Relatedly, it would have been interesting to involve judges directly in IVR-based rating sessions, that is, having the judges explore and evaluate the VIVA task productions within the "ecological environments" in which such works were generated (i.e., the *tilt brush* environment). Such procedures would of course be more taxing for the raters and for the study's logistics, but it could have served the purpose to cross-validate our mere evaluation of a 2D screenshot – losing much of the features of the original productions – with the *in-situ* ratings of productions as 3D objects.[5] Although it is established that IVR increases attention to details over 2D (Schöne, Sylvester, Radtke, & Gruber, 2020), creativity and design research regularly uses 2D representations of 3D objects for judgments of products (Wojtczuk & Bonnardel, 2011).

The current findings regarding the psychometric properties of the VIVA's scores (as well as the selected models) are somewhat limited by the sample size used in this study. Yet, evidences from various analytical approaches used in this study converge toward the same pattern of encouraging findings with regard to the structural validity, internal reliability, and criterion validity of the VIVA's scores. Although this study remains an initial step toward the development of psychometric paradigms for the measurement of creativity in IVR, this limitation requires further investigations. In particular, the generalizability of the findings derived from the IRT models should be further explored, as these models are heavily parameterized given the observed sample size in our study. Besides the fact that psychometric properties are attributes of the scores and not the tests themselves (i.e., psychometric properties are sample-specific), a researcher reusing our analytical approach may also obtain different psychometric properties from using different judges or different products to judge. Therefore, we encourage the replication of our findings with different products and different judges in larger samples to establish the generalizability of the findings obtained in this study. Finally, this work would benefit from the assessment of more psychometric properties such as delayed alternate form reliability (essential for application in longitudinal studies; Barbot, 2019), discriminant validity with more dimensions of the productions (e.g., technical quality), criterion validity with more external measures (in particular, graphic tasks not implemented in IVR), or predictive validity of "real-life" creative behaviors.

Future directions

Despite limitations, this study offers initial evidences that may pave the way for an important new era of work focused on leveraging creativity in virtual environments. In a nutshell, the study has established that the proposed assessment paradigm works reasonably well for its intended purpose. Augmenting pioneering findings (Ward & Sonneborn, 2011), it also shows that IVR users are able to "bring" their creative potential and technical skills in virtual reality settings. This conclusion is particularly important because, while we have essentially focused on a "product" perspective of creativity, the IVR setting offers a formidable opportunity to extend this scope to other dimensions of creativity. Indeed, IVR technology is particularly suited to design and implement assessments that bridge together the "4-Ps" of creativity (Rhodes, 1961), simultaneously capturing (1) individual differences (the *Person*), (2) track closely the stages involved in bringing to life the creative production (the *Process*), (3) dynamically manipulating environmental cues that can stimulate or inhere creativity (the *Press*), and (4) yield a digital output (i.e., the *Product)* which, as demonstrated here, can be further assessed using the classic CAT paradigm (Amabile, 1996) and modern extensions (Myszkowski & Storme, 2019a).

While these perspectives are only scratching the surface of what IVR has to offer to the realm of creativity (assessment), the practical and societal impact of such application are tremendous. First, in the current context of worldwide pandemic, confinement and social isolation, providing IVR-based solutions for assessment, education, treatment and rehabilitation focusing on creativity are ever more necessary. Assessment of creativity in IVR has proved feasible in the present study, and such IVR-based assessment could be applied on a large-scale across the world with consistent methods, prompts, and settings, making test-performance more comparable across individuals (Blascovich et al., 2002). This goal seems to be – with or without pandemic – a general direction in the field of assessment. Furthermore, it can be extended in a straightforward manner to other creativity assessment paradigms, including divergent thinking tasks (such as simulated Alternative Use Tasks) or block building production tasks in environements such as VRBox (Fröhlich, Alexandrovsky, Stabbert, Döring, & Malaka, 2018).

Ideally, measures of ability and educational achievement will incorporate these developments: illustrated by the 4P approach of creativity assessment in IVR outlined above, imagine how much more information we might glean from an IVR integrative assessment that replaces or supplements the classic SATs/ACTs tests. Further, this becomes a prime time to include creativity as part of such an assessment. Incorporating creativity into admissions assessments can offer many benefits. In addition to providing a fuller picture of the student and (at least indirectly) encouraging more creative instruction in the classroom because it is on the test (Beghetto, Kaufman, & Baer, 2014), it also can enhance diversity and equity in admissions because there are virtually no differences by culture or ethnicity on creativity tests, whereas such differences exist on ability/achievement test measures (Kaufman, 2010, 2015; Luria & Kaufman, 2017). In all, assessing creativity in IVR is only a first, but necessary step to increasingly engage people with creativity in digital realities. Only with such assessment will we be able to develop evidence-based and tailored programs that stimulate people's creativity, and corresponding individual and societal benefits.

Notes

1. In contrast to the experimental setting, the preview featured only seating experience using the Oculus Rift.
2. The order of both alternate forms was counterbalanced.
3. Across all the ratings, there was only 1 missing rating (of aesthetic quality) for 1 product. This observation was skipped in investigations of model fit and reliability.
4. Rules of thumb and simulations in the SEM literature (e.g., Wolf et al., 2013) suggest that such simple models does not require more than 60 participants for a stable estimation.
5. As a matter of fact, tilt brush allows exporting room-scale creations in a range of formats (e.g., .fbx, .usd, . json) which could let raters evaluate these productions as 3D object whether in IVR or not.

Acknowledgments

The preparation of this article was supported by grant RFP-15-05 to Baptiste Barbot from the Imagination Institute (www.imagination-institute.org), funded by the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the view of the Imagination Institute or the John Templeton Foundation. We thank Alexandra Blanchard, Brianna Heuser, Rosemary Leo, Catalina Mourgues, Danielle Saliani, and Mansi Shah for help with data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Availability of data and material

Available from the corresponding author upon reasonable request.

Funding

The preparation of this article was supported by grant RFP-15-05 to Baptiste Barbot from the Imagination Institute (www.imagination-institute.org), funded by the John Templeton Foundation.

ORCID

Baptiste Barbot http://orcid.org/0000-0002-5096-2596

References

Amabile, T. M. (1996). *Creativity in context: update to 'the social psychology of creativity'*. Vol. xviii. Westview Press. http://apa.org/psycinfo/1996-97996-000

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. doi:10.1007/BF02293814

Avitia, M. J., & Kaufman, J. C. (2014). Beyond g and c: the relationship of rated creativity to long-term storage and retrieval (Glr). *Psychology of Aesthetics, Creativity, and the Arts*, *8*(3), 293. doi:10.1037/a0036772

Baer, J., & Kaufman, J. C. (2005). Bridging generality and specificity: the amusement park theoretical (APT) model of creativity. *Roeper Review*, *27*(3), 158–163. doi:10.1080/02783190509554310

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. doi:10.1207/s15326934crj1601_11

Baer, J., Kaufman, J. C., & Riggs, M. (2009). Brief report: rater-domain interactions in the consensual assessment technique. *The International Journal of Creativity & Problem Solving*, *19*(2), 87–92.

Barbot, B. (2019). Measuring creativity change and development. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 203–210. doi:10.1037/aca0000232

Barbot, B. (2020). Creativity and self-esteem in adolescence: a study of their domain-specific, multivariate relationships. *The Journal of Creative Behavior*, *54*(2), 279–292. doi:10.1002/jocb.365

Barbot, B. (2021). The intertwined development of identity and creativity: immersing in the digital self. In S. W. Russ, J. Hoffmann, & J. C. Kaufman (Eds.), *The Cambridge handbook of lifespan development of creativity (Cambridge handbooks in psychology)* (pp. 159–175). Cambridge University Press.

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

Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (re) setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 233–240. doi:10.1037/aca0000233

Barbot, B., & Heuser, B. (2017). Creativity and identity formation in adolescence: A developmental perspective. In *The creative self: effect of beliefs, self-efficacy, mindset, and identity* (pp. 87–98). Elsevier Academic Press. doi:10.1016/B978-0-12-809790-8.00005-4

Barbot, B., & Kaufman, J. C. (2020). What makes immersive virtual reality the ultimate empathy machine? Discerning the underlying mechanisms of change. *Computers in Human Behavior*, *111*, 106431. doi:10.1016/j.chb.2020.106431

Barbot, B., & Lubart, T. (2012). Creative thinking in music: its nature and assessment through musical exploratory behaviors. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(3), 231–242. doi:10.1037/a0027307

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

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*(3), 209–223. doi:10.1016/j.tsc.2012.04.006

Beghetto, R. A., Kaufman, J. C., & Baer, J. (2014). *Teaching for creativity in the common core classroom*. Teachers College Press.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103–124. doi:10.1207/S15327965PLI1302_01

Brivio, E., Serino, S., Negro Cousa, E., Zini, A., Riva, G., & De Leo, G. (2020). Virtual reality and 360° panorama technology: A media comparison to study changes in sense of presence, anxiety, and positive emotions. *Virtual Reality*, *25*(2), 303–311. doi:10.1007/s10055-020-00453-7

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *The British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. doi:10.1111/j.2044-8317.2012.02050.x

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

Chang, Y., Kao, J. -Y., & Wang, Y. -Y. (2022). Influences of virtual reality on design creativity and design thinking.

Thinking Skills and Creativity, *46*, 101127. doi:10.1016/j.tsc.2022.101127

Chen, Y. -C., Chang, Y. -S., & Chuang, M. -J. (2022). Virtual reality application influences cognitive load-mediated creativity components and creative performance in engineering design. *Journal of Computer Assisted Learning*, *38*(1), 6–18. doi:10.1111/jcal.12588

Child, I. L. (1964). Observations on the meaning of some measures of esthetic sensitivity. *The Journal of Psychology*, *57*(1), 49–64. doi:10.1080/00223980.1964.9916671

Cropley, D. H. (2015). Teaching engineers to think creatively. In R. Wegerif, L. Li, & J. C. Kaufman (Eds.), *The International Handbook of Research on Teaching Thinking* (pp. 402–410). Routledge.

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 159. doi:10.1037/aca0000220

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: A review of research in virtual reality. *Frontiers in Psychology*, *6*, 6. doi:10.3389/fpsyg.2015.00026

Faria, A. L., Andrade, A., Soares, L., & I Badia, S. B. (2016). Benefits of virtual reality based cognitive rehabilitation through simulated activities of daily living: A randomized controlled trial with stroke patients. *Journal of Neuroengineering and Rehabilitation*, *13*(1), 96. doi:10.1186/s12984-016-0204-z

Florida, R. (2014). The creative class and economic development. *Economic Development Quarterly*, *28*(3), 196–205. doi:10.1177/0891242414541693

Forgeard, M. J. (2013). Perceiving benefits after adversity: the relationship between self-reported posttraumatic growth and creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(3), 245. doi:10.1037/a0031223

Forgeard, M. J., & Kaufman, J. C. (2016). Who cares about imagination, creativity, and innovation, and why? A review. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(3), 250. doi:10.1037/aca0000042

Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological Medicine*, *47*(14), 1–8. doi:10.1017/S003329171700040X

Fröhlich, T., Alexandrovsky, D., Stabbert, T., Döring, T., & Malaka, R. (2018). VRBox: a virtual reality augmented sandbox for immersive playfulness, creativity and exploration. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 153–162. 10.1145/3242671.3242697

Fürst, G. (2020). Measuring creativity with planned missing data. *The Journal of Creative Behavior*, *54*(1), 150–164. doi:10.1002/jocb.352

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

Goncalo, J. A., Vincent, L. C., & Krause, V. (2015). The liberating consequences of creative work: how a creative outlet lifts the physical burden of secrecy. *Journal of Experimental Social Psychology*, *59*, 32–39. doi:10.1016/j.jesp.2015.03.004

Götz, K. O. (1985). *VAST: visual aesthetic sensitivity test* (4th ed.). Concept Verlag.

Graessler, I., & Taplick, P. (2019). Supporting creativity with virtual reality technology. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 2011–2020. doi:10.1017/dsi.2019.207

Grigorenko, E. L. (2019). Creativity in digital reality/creatividad en la realidad digital. *Studies in Psychology*, *40*(3), 585–607. doi:10.1080/02109395.2019.1660122

Groyecka-Bernard, A., Karwowski, M., & Sorokowski, P. (2021). Creative thinking components as tools for reducing prejudice: evidence from experimental studies on adolescents. *Thinking Skills and Creativity*, *39*, 100779. doi:10.1016/j.tsc.2020.100779

Guan, J. -Q., Wang, L. -H., Chen, Q., Jin, K., & Hwang, G. -J. (2021). Effects of a virtual reality-based pottery making approach on junior high school students' creativity and learning engagement. *Interactive Learning Environments*, *0*(0), 1–17. doi:10.1080/10494820.2021.1871631

Guilford, J. P. (1950). Creativity. *The American Psychologist*, *5* (9), 444–454. doi:10.1037/h0063487

Hennessey, B. A., & Amabile, T. M. (2009). Creativity. *Annual Review of Psychology*, *61*(1), 569–598. doi:10.1146/annurev.psych.093008.100416

Hoffmann, J., Ivcevic, Z., & Brackett, M. (2016). Creativity in the age of technology: measuring the digital creativity of millennials. *Creativity Research Journal*, *28*(2), 149–153. doi:10.1080/10400419.2016.1162515

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., . . . Beaty, R. E. (2022). Divergent semantic integration (DSI): extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*. doi:10.3758/s13428-022-01986-2

Kapoor, H., & Kaufman, J. C. (2020). Meaning-making through creativity during COVID-19. *Frontiers in Psychology*, *11*. doi:10.3389/fpsyg.2020.595990

Kaufman, J. C. (2010). Using creativity to reduce ethnic bias in college admissions. *Review of General Psychology*, *14*(3), 189. doi:10.1037/a0020133

Kaufman, J. C. (2015). Why creativity isn't in IQ tests, why it matters, and why it won't change anytime soon probably. *Journal of Intelligence*, *3*(3), 59–72. doi:10.3390/jintelligence3030059

Kaufman, J. C. (2018). Finding meaning with creativity in the past, present, and future. *Perspectives on Psychological Science*, *13*(6), 734–749. doi:10.1177/1745691618771981

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

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

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. doi: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. doi:10.1037/a0034809

Kaufman, S. B., Kozbelt, A., Silvia, P., Kaufman, J. C., Ramesh, S., & Feist, G. J. (2016). Who finds Bill Gates sexy? Creative mate preferences as a function of cognitive ability, personality, and creative achievement. *The Journal of Creative Behavior*, *50*(4), 294–307. doi:10.1002/jocb.78

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*(2), 96–106. doi:10.1016/j.tsc.2007.04.002

Kozbelt, A., Seidel, A., ElBassiouny, A., Mark, Y., & Owen, D. R. (2010). Visual selection contributes to artists' advantages in realistic drawing. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(2), 93–102. doi:10.1037/a0017657

Lau, K. W., & Lee, P. Y. (2015). The use of virtual reality for creating unusual environmental stimulation to motivate students to explore creative ideas. *Interactive Learning Environments*, *23*(1), 3–18. doi:10.1080/10494820.2012.745426

Li, H., Du, X., Ma, H., Wang, Z., Li, Y., & Wu, J. (2022). The effect of virtual-reality-based restorative environments on creativity. *International Journal of Environmental Research and Public Health*, *19*(19), 19. doi:10.3390/ijerph191912083

Liu, C. -W., & Chalmers, R. P. (2018). Fitting item response unfolding models to likert-scale data using mirt in R. *PLoS One*, *13*(5), 1–22. doi:10.1371/journal.pone.0196292

Luria, S. R., & Kaufman, J. C. (2017). Examining the relationship between creativity and equitable thinking in schools. *Psychology in the Schools*, *54*(10), 1279–1284. doi:10.1002/pits.22076

Maguire, L. E., Wanschura, P. B., Battaglia, M. M., Howell, S. N., & Flinn, J. M. (2015). Participation in active singing leads to cognitive improvements in individuals with dementia. *Journal of the American Geriatrics Society*, *63*(4), 815–816. doi:10.1111/jgs.13366

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Mattila, O., Korhonen, A., Pöyry, E., Hauru, K., Holopainen, J., & Parvinen, P. (2020). Restoration in a virtual reality forest environment. *Computers in Human Behavior*, *107*, 106295. doi:10.1016/j.chb.2020.106295

Meier, M., Unternaehrer, E., Schorpp, S. M., Wenzel, M., Benz, A., Bentele, U. U., . . . Prüssner, J. C. (2020). The opposite of stress: the relationship between vagal tone, creativity, and divergent thinking. *Experimental Psychology*, *67*(2), 150–159. doi:10.1027/1618-3169/a000483

Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, *14*(1), 59–71. doi:10.1177/014662169001400106

Muraki, E., & Muraki, M. (2016). Generalized partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: models* (1st ed., pp. 127–137). Chapman and Hall/CRC.

Myszkowski, N. (2019a). The first glance is the weakest: "Tasteful" individuals are slower to judge visual art. *Personality and Individual Differences*, *141*, 188–195. doi:10.1016/j.paid.2019.01.010

Myszkowski, N. (2019b). Development of the R library 'jrt': automated item response theory procedures for judgment data and their application with the consensual assessment technique. *Psychology of Aesthetics, Creativity, and the Arts*. doi:10.1037/aca0000287

Myszkowski, N. (2020). Aesthetic sensitivity. In M. Nadal & O. Vartanian (Eds.), *The oxford handbook of empirical aesthetics*. Oxford University Press. doi:10.1093/oxfordhb/9780198824350.013.40

Myszkowski, N., Çelik, P., & Storme, M. (2020). Commentary on Corradi et al.'s (2019) new conception of aesthetic sensitivity: is the ability conception dead? *British Journal of Psychology*, *111*(4), 659–662. doi:10.1111/bjop.12440

Myszkowski, N., & Storme, M. (2017). Measuring "Good Taste" with the visual aesthetic sensitivity test-revised (VAST-R). *Personality and Individual Differences*, *117*, 91–100. doi:10.1016/j.paid.2017.05.041

Myszkowski, N., & Storme, M. (2019a). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 167. doi:10.1037/aca0000225

Myszkowski, N., & Storme, M. (2019b). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 167–175. doi:10.1037/aca0000225

Myszkowski, N., Storme, M., & Zenasni, F. (2016). Order in complexity: how Hans Eysenck brought differential psychology and aesthetics together. *Personality and Individual Differences*, *103*, 156–162. doi:10.1016/j.paid.2016.04.034

Myszkowski, N., Storme, M., Zenasni, F., & Lubart, T. (2014). Is visual aesthetic sensitivity independent from intelligence, personality and creativity? *Personality and Individual Differences*, *59*, 16–20. doi:10.1016/j.paid.2013.10.021

Myszkowski, N., & Zenasni, F. (2016). Individual differences in aesthetic ability: the case for an aesthetic quotient. *Frontiers in Psychology*, *7*(750). doi:10.3389/fpsyg.2016.00750

Nelson, J., & Guegan, J. (2019). "I'd like to be under the sea": contextual cues in virtual environments influence the orientation of idea generation. *Computers in Human Behavior*, *90*, 93–102. doi:10.1016/j.chb.2018.08.001

Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models* (1 ed.). Routledge.

Obeid, S., & Demirkan, H. (2020). The influence of virtual reality on design process creativity in basic design studios. *Interactive Learning Environments*, *0*(0), 1–19. doi:10.1080/10494820.2020.1858116

Patz, R., Grzymala, M., & Gengnagel, C. (2019). Bodily design processes in immersive virtual environments. *Design Modelling Symposium Berlin*, 350–359.

Pelowski, M., Leder, H., & Tinio, P. P. (2017). Creativity in the visual arts. In J. C. Kaufman, V. P. Glăveanu, & J. Baer (Eds.), *The Cambridge handbook of creativity across domains* (pp. 80–109). Cambridge University Press.

Peña, J., & Blackburn, K. (2013). The priming effects of virtual environments on interpersonal perceptions and behaviors. *Journal of Communication*, *63*(4), 703–720. doi:10.1111/jcom.12043

Pennebaker, J. W., & Seagal, J. D. (1999). Forming a story: the health benefits of narrative. *Journal of Clinical Psychology*, *55*(10), 1243–1254. doi:10.1002/(SICI)1097-4679(199910)55:10<1243:AID-JCLP6>3.0.CO;2-N

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, *39*(2), 83–96. doi:10.1207/s15326985ep3902_1

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology & Marketing*, *26*(5), 470–478. doi:10.1002/mar.20283

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*(2), 176–186. doi:10.1037/aca0000230

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

Ritter, S. M., Damian, R. I., Simonton, D. K., van Baaren, R. B., Strick, M., Derks, J., & Dijksterhuis, A. (2012). Diversifying experiences enhance cognitive flexibility. *Journal of Experimental Social Psychology*, *48*(4), 961–964. doi:10.1016/j.jesp.2012.02.009

Roberts, R. O., Cha, R. H., Mielke, M. M., Geda, Y. E., Boeve, B. F., Machulda, M. M., . . . Petersen, R. C. (2015). Risk and protective factors for cognitive impairment in persons aged 85 years and older. *Neurology*, *84*(18), 1854–1861. doi:10.1212/WNL.0000000000001537

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101–139.

Rosenberg, R. S., Baughman, S. L., & Bailenson, J. N. (2013). Virtual superheroes: using superpowers in virtual reality to encourage prosocial behavior. *PLoS One*, *8*(1), e55003. doi:10.1371/journal.pone.0055003

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). doi:10.18637/jss.v048.i02

Runco, M. A., & Abdullah, A. M. (2014). Why isn't creativity being supported? Distressing analyses of grants and awards for creativity research–or lack thereof. *Creativity Research Journal*, *26*(2), 248–250. doi:10.1080/10400419.2014.901100

Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., Van den Noortgate, W., & Barbot, B. (2020). Does the fourth-grade slump in creativity actually exist? A meta-analysis of the development of divergent thinking in school-age children and adolescents. *Educational Psychology Review*, *33*(1), 275–298. doi:10.1007/s10648-020-09547-9

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

Schöne, B., Sylvester, R. S., Radtke, E. L., & Gruber, T. (2020). Sustained inattentional blindness in virtual reality and under conventional laboratory conditions. *Virtual Reality*, *25*(1), 209–216. doi:10.1007/s10055-020-00450-w

Sica, L. S., Ragozini, G., Palma, T. D., & Sestito, L. A. (2019). Creativity as identity skill? Late adolescents' management of identity, complexity and risk-taking. *The Journal of Creative Behavior*, *53*(4), 457–471. doi:10.1002/jocb.221

Simonton, D. K. (2009). *Genius 101*. Springer Publishing Company.

Simonton, D. K. (2012). Taking the US patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal*, *24*(2–3), 97–106. doi:10.1080/10400419.2012.676974

Tang, C., Mao, S., Naumann, S. E., & Xing, Z. (2022). Improving student creativity through digital technology products: A literature review. *Thinking Skills and Creativity*, *44*, 101032. doi:10.1016/j.tsc.2022.101032

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*(3), 73–90. doi:10.3390/jintelligence3030073

Taylor, C. L., & Barbot, B. (2021). Dual pathways in creative writing processes. *Psychology of Aesthetics, Creativity, and the Arts*. No Pagination Specified-No Pagination Specified. 10.1037/aca0000415

Taylor, C. L., Kaufman, J. C., & Barbot, B. (2021). Measuring creative writing with the storyboard task: the role of effort and story length. *The Journal of Creative Behavior*, *55*(2), 476–488. doi:10.1002/jocb.467

Tennant, M., McGillivray, J., Youssef, G. J., McCarthy, M. C., & Clark, T. -J. (2020). Feasibility, acceptability, and clinical implementation of an immersive virtual reality intervention to address psychological well-being in children and adolescents with cancer. *Journal of Pediatric Oncology Nursing*, *37*(4), 265–277. doi:10.1177/1043454220917859

Thornhill-Miller, B., & Dupont, J. -M. (2016). Virtual reality and the enhancement of creativity and innovation: under recognized potential among converging technologies? *Journal of Cognitive Education and Psychology*, *15*(1), 102–121. doi:10.1891/1945-8959.15.1.102

Walberg, H. J. (1988). Creativity and talent as learning. In R. J. Sternberg (Ed.), *The Nature of Creativity: Contemporary Psychological Perspectives* (pp. 340–361). Cambridge University Press.

Wang, Y. -Y., Weng, T. -H., Tsai, I. -F., Kao, J. -Y., & Chang, Y. -S. (in press). Effects of virtual reality on creativity performance and perceived immersion: A study of brain waves. *British Journal of Educational Technology*. *n/a* (n/a. doi: 10.1111/bjet.13264

Ward, T., & Sonneborn, M. (2009). Creative expression in virtual worlds: imitation, imagination, and individualized collaboration. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(4), 211–221. doi:10.1037/a0016297

Ward, T. B., & Sonneborn, M. S. (2011). Creative expression in virtual worlds: imitation, imagination, and individualized collaboration. *Psychology of Popular Media Culture*, *1*(S), 32–47. doi:10.1037/2160-4134.1.S.32

Wojtczuk, A., & Bonnardel, N. (2011). Designing and assessing everyday objects: impact of externalisation tools and judges' backgrounds. *Interacting with Computers*, *23*(4), 337–345. doi:10.1016/j.intcom.2011.05.004

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. doi:10.1177/0013164413495237

Yang, X., Lin, L., Cheng, P. -Y., Yang, X., Ren, Y., & Huang, Y. -M. (2018). Examining creativity through a virtual reality support system. *Educational Technology Research and Development*, *66*(5), 1231–1254. doi:10.1007/s11423-018-9604-z

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, *51*(2), 879–894. doi:10.3758/s13428-018-1137-1