Giving the Art Greater Weight in Art Psychology: RizbA, a Quantitative Questionnaire for Formal Picture Analysis

Kerstin Schoch

University of Applied Sciences and Arts Ottersberg, Germany
Witten/ Herdecke University, Germany

E-mail address: kerstin.schoch@hks-ottersberg.de

Thomas Ostermann

Witten/ Herdecke University, Germany

E-mail address: thomas.ostermann@uni-wh.de

ARTICLE INFO

Keywords:

Contemporary art
Formal picture analysis
Psychometrics
Questionnaire
Visual art

Article history:

Received 09 September 2020
Received in revised form 12 November 2020
Accepted 22 November 2020

ISSN: 2354-0036

DOI: 10.2478/ctra-2020-0019

ABSTRACT

In empirical art psychology and creativity research most studies focus on the psychological correlates of art. Only few go beyond treating artworks as categorical data (e.g. abstract vs. representational) and consider artworks in detail. In part this is due to the lack of reliable quantitative measurements. The rating instrument for two-dimensional pictorial works (*RizbA*) makes a difference to current research designs. The current study validates the questionnaire on a representative sample of contemporary visual art, consisting of 318 images depicting works by artists from different cultural areas dated to the 21st century. In a randomized test-retest design, the pictorial material was rated by 506 (T1) and 238 (T2) art experts using *RizbA*. Statistical quality criteria, such as item difficulty, capacity of differentiation, test-retest reliability, and intraclass correlation were calculated. Principal component analysis (PCA) and indices of factor similarity were computed. The overall test's capacity for differentiation yields partial eta-squared of .31 (T1) and .40 (T2). Test-retest reliability is .86. PCA reveals an eight-factor solution, which is largely consistent across both measurement points. Tucker's coefficient of congruence ranges between |.71| and |1.00|. Intraclass correlation coefficients are .86 (T1) and .73 (T2). This study indicates generalizability of the questionnaire to contemporary artworks. Although a conclusion on the factors' structure cannot be drawn yet, results are very promising. As the first reliable quantitative tool for formal picture analysis, *RizbA* allows more detailed examination of visual art and its psychological correlates. This broadens research methodology by giving art greater weight in art psychology and creativity research.

* The authors gratefully acknowledge Thomas Gengenbach for providing theoretical and practical support regarding IT, implementation of the online survey, and data structuring. Many thanks to Ina Blümel for being a great mentor in conducting this study as part of the Open Science Fellows Program. Thanks also to Rebecca Kahn for the feedback on the manuscript. This study was funded by the Open Science Fellows Program by Wikimedia Germany, Stifterverband and VolkswagenStiftung.

INTRODUCTION

Although psychological research on art has some tradition, artworks themselves, including their formal content, have received surprisingly little attention in quantitative research. Although Hayn-Leichsenring, Kenett, Schulz, and Chatterjee (2020) state that it is not useful to treat abstract paintings as a single category, art is usually operationalized in terms of a categorical independent variable. In fact, most studies in art psychology, creativity research, and empirical aesthetics focus on art in general and its psychological correlates, while only few consider artworks in detail or at all. For example, in experiments, visual art is often operationalized in terms of broad, sometimes even dichotomous categories, such as art vs. non-art (e.g., Hager, Hagemann, Danner, & Schankin, 2012), genre (Commare, Rosenberg, & Leder, 2018), epoch of art (e.g., Leder, Gerger, Dressler, & Schabmann, 2012), representational vs. abstract (e.g., Gridley, 2013), degree of abstraction (e.g., Specht, 2007), symmetry vs. asymmetry (e.g., Leder et al., 2018) or beauty vs. ugliness (e.g., Ishizu & Zeki, 2011). Only a few studies go into further detail; for example, by identifying different forms of complexity in artworks (Nadal, Munar, Marty, & Cela-Conde, 2010) or by examining the relationship between perception of image complexity and content-related processes in paintings (Commare et al., 2018). Beyond that, hardly any study uses contemporary artworks as pictorial stimuli. Still, these approaches only partially consider art with all its various layers and the sense-making processes that might be related to it: embodied, felt, emotional, intellectual, and intersubjective (Sullivan & McCarthy, 2009). Treating artworks categorically yields only a limited range of information (MacLagan, 1999) of the content contained in pictures.

As discussed by Schoch, Gruber, and Ostermann (2017) and Schoch and Ostermann (under review) there is a severe lack of instruments for a detailed analysis of artworks, that are also quantitative, standardized, and reliable, allow inference statistics, and are validated on and meant to be applicable to all sorts of two-dimensional pictorial works. Only one tool stemming from a neuropsychological perspective shares a similar approach with this work: *The Assessment of Art Attributes* (Chatterjee, Widick, Sternschein, Smith, & Bromberger, 2010; Penn Center for Neuroaesthetics, 2019). By connecting art and science, it aims to quantify artworks including to both formal-perceptual and content-representational attributes. The scale consists of 14 items assessing broad categories. The results suggest medium to high agreement between raters, while using non-parametric measures of correlation and having previously conducted a rater training with training slides regarding the attributes. Unfortunately, other statistical quality criteria, such as the test's capacity to differentiate between pictures or subgroups of pictures or factors analysis, are not reported. Moreover, it has been tested on a small sample of prominent

paintings from the western canon, but as far as we can tell, not on other pictorial material, and it was not further validated.

Consequently, more detailed validated approaches are needed to measure artworks in nuanced profiles to fill this gap. If genuine aspects of artworks are left out of research methodology, this is due in part to a lack of reliable quantitative measures available for this purpose. This is where the rating instrument for two-dimensional pictorial works (*RizbA*) contributes to the current lines of research. In a first pilot study (Schoch et al., 2017), the questionnaire was developed and empirically tested on a small heterogeneous sample of pictorial works created by non-professionals. This resulted in a preliminary test version and encouraging statistical results regarding reliability. In a second study (Schoch & Ostermann, under review) *RizbA* was validated on a large representative sample of amateur pictorial works and a large sample of experts rating those, yielding moderate to high reliabilities and allowing generalizability to non-professional art.

METHOD

Since the two previous studies successfully validated *RizbA* on amateur pictorial works, the current study aims to extend the area of application to professional visual art. It validates the instrument in terms of its generalizability to contemporary artworks. A large representative sample of pictorial works is used, created by artists from different cultural areas and dated to the 21st century.

The Rating Instrument for Two-Dimensional Pictorial Works (*RizbA*)

RizbA is a 26-item questionnaire referring to formal picture analysis (Streb, 1984; Stuhler-Bauer & Elbing, 2003), originating from the humanities and the arts and focusing on maximization of objectivity. It focuses on structuring art experts' ratings and includes formal aspects such as representation, color, shape, motion, composition, and expression. The test was developed in a previous pilot study (Schoch et al., 2017) and successfully validated in a second study with a large representative sample (Schoch & Ostermann, under review), both using pictorial works created by amateurs. The questionnaire uses a bipolar, six-point Likert-scale, which is discretely scaled and verbally anchored in shades of agreement (0 = *strongly disagree*, 5 = *strongly agree*). Using the questionnaire, raters receive a brief instruction to rate the image presented using the *RizbA* questionnaire. They are asked to focus on pre-dominant overall expression of a picture rather than any single details, while endorsing that there is no right or wrong, but only their valuation.

Pictorial Material.

The image material consists of artworks by contemporary artists from different cultural areas, including Europe, Asia, Africa, Northern and Southern America. We chose contemporary art for two reasons. Firstly, although art has evolved in the last centuries, it is

surprising that empirical research with contemporaries still uses only a small range of hundred-year-old, predominantly western images, such as impressionist or expressionist paintings when it comes to artistic stimulus material. One practical reason for this might be legal aspects of image rights. Fortunately, with digitization and Open GLAM initiatives (GLAM Wiki, 2020) availability of art has currently considerably changed. We thus took this chance to use contemporary artworks for a contemporary study. The second reason for using contemporary works is comparability, since all previous validation studies (Schoch et al., 2017; Schoch & Ostermann, under review) and pilot studies (Epstein, 2019; Janßen, 2018) on *RizbA* used amateur works from the 21st century.

Data was gathered using *prometheus*, which is a distributed digital image archive connecting databases from various institutes, research facilities, and museums (<https://www.prometheus-bildarchiv.de>; Simon & Verstegen, 2004). Databases not relevant to the subject were excluded from the search engine (e.g., architecture, archeology, early and prehistory, Egyptology, churches and castles, graphic prints, video). The archive was searched using the advanced search option starting with image material dated 2018 and going backwards chronologically. Hits were further selected based on additional criteria the first of which being that the year of their making had to be in the 21st century.

Second, only drawings, paintings, collages, and mixed-techniques in between were used, since the test has not yet been validated for other than handmade two-dimensional works. Thus, machined works (e.g., photography), three-dimensional work (e.g., sculpture) and printing techniques (e.g., etching) were excluded. Since they make use of more elaborated tools, such as a photo camera or a printing press, those techniques entail different creation processes than using a pencil or brush, where the ductus of the creator is directly recognizable. It is highly likely that *RizbA* also works for non-handmade techniques, but it needs further testing and validation to have empirical evidence considering that.

Third, the image had to comprise a whole view. Images showing only a detail of an artwork or a whole exhibition space displaying the artwork were excluded. Fourth, images had to offer a sufficiently good picture quality (e.g., resolution, sharpness, picture section). The resulting sample was reduced to one image per artist, choosing the most recent piece.

The required sample size of images was calculated based on existing guidelines on principal component analysis (PCA). We oriented towards Guadagnoli and Velicer (1988) as well as Osborne and Costello (2009), in which a subject to item ratio of at least 10:1 is proposed. As a consequence, a minimum sample size of at least 260 images was calculated to be sufficient for the study.

The systematic search resulted in a final sample of 318 images of artworks by professional artists differing in gender, ethnicity, and origin dated from 2005 to 2018 (see Table 1). Since *prometheus* is a distributed archive with data stemming from different institutions, the digital image size varied between 18 kilobytes and 7.4 megabytes. Meeting the requirements of the online survey provider, larger images were reduced to a maximum of 640 kilobytes by using Apple Photos software.

Table 1
Pictorial material: Dating of artworks

Dating (year)	Number of artworks selected
2018	0
2017	0
2016	3
2015	19
2014	35
2013	20
2012	22
2011	14
2010	21
2009	25
2008	22
2007	37
2006	30
2005	70
In total	318

The Study

Study Design

An online study design was chosen using *SoSci Survey* (Leiner, 2018) with additional use of PHP elements. The study consisted of a test (T1) and a retest (T2) at intervals of a minimum of 14 days (see Figure 1), in which experts were asked to rate the image sample using *RizbA*. Each rater was presented the same six images at both times of measurement. Raters with a professional training in visual art were recruited via email, websites, and social media. For this purpose, 122 universities, 48 associations, 23 art museums and 742 individuals in Germany, Austria, and Switzerland were contacted inviting all their professional staff, students, and associates to participate and disseminate the call for raters among others. No rater training was conducted. As an incentive, vouchers were raffled among participants.

Figure 1. Study design: Test-retest design

Raters

As a criterion for inclusion to ensure raters know the art terms used, only professionals who have an academic degree related to visual art or are students who have been studying an art-related subject for at least one year were requested to participate. Only relevant disciplines were allowed. These included art history, art pedagogy, visual art, art therapy, design, graphic design, art sciences, image sciences, restoration and others.

There are only a few empirically based guidelines regarding the statistically required sample size of raters. Some literature recommends two or three observations per subject, given $\alpha = .05$ and a required reliability of at least .40 (Walter, Eliasziw, & Donner, 1998). Since this is not an application study, but the constitutive calibration of a new questionnaire, either way we opted for a larger number of raters. Addressing sample size and ICCs, for an exploratory approach, an *a priori* set cut-off point and a pilot study are recommended (LeBreton & Senter, 2008). The first pilot study (Schoch et al., 2017) yielded moderate to very strong rater agreement with eight to twelve raters per image. Since both preceding studies (Schoch et al., 2017; Schoch & Ostermann, under review) resulted in sufficiently high inter-rater agreement and test-retest reliability, the current investigation concentrates on artistic stimulus material rather than on the raters' judgement. This is why the current one aims for only ten raters per picture at T1, in comparison to the previous study aiming for the double number.

Procedure

In the online survey, participants were introduced to using the *RizbA* scale and asked to rate each picture consecutively presented to them using all 26 items. For each rater, six images were randomly, successively drawn without returning them. When all images of the entire picture sample had been drawn from the hat and rated once, the process started again. The order in which items were presented was randomized between subjects.

Fourteen days after completing the first survey (T1), raters were electronically invited to the retest, at which they were shown the same selection of pictures as in the first

part and again rated them using the *RizbA* scale. Raters could not skip items, but had to answer all questions to proceed. At the end of both tests, raters were asked whether they had completed the questionnaire seriously, outlining that the participation is anonymous and there will be no consequences if they didn't. In T1, only datasets of raters who finished rating all six images were used in the analysis. Raters who stated not to have completed the questionnaire seriously were not invited to the retest (T2). For each rater, a code was automatically generated at the end of T1, maintaining anonymity and enabling later assignment from T1 and T2 responses. A database was automatically established, in which information was stored.

Statistical Analysis

Only complete datasets were used, in which all six presented images were fully processed. Data from prematurely terminated surveys was excluded. For reasons of comparability, statistical analyses were conducted analogically to the recent validation study on amateur works.

Analysis of the questionnaire's quality criteria focused on two units. For calculation of item means, item standard deviation, item difficulty, capacity of differentiation, test-retest reliability, factor analysis, and index of factor similarity pictorial works were the unit of analysis. For inter-rater reliability, raters served as the relevant unit. The data included three levels (items, pictures, raters), of which the relevant one was chosen for each analysis, while the other two were averaged. Because each rater assessed only six images from the entire pictorial material, on a rater level the dataset inherently included a large amount of empty cells. For analyses regarding raters' agreement, subjects were therefore merged to new generic combined raters. Due to this restructuring of the data and different numbers of ratings per image, missing values occurred, which were not imputed. For analysis, only those generic raters with less than 20 % missing values were used, so as not to lose too much data through listwise deletion. This cut-off point was defined *a posteriori*, based on the descriptive statistics. For structuring the datasets, Excel including Visual Basic for Applications was used. All statistical analyses were analogically conducted for both times of measurements, test (T1) and retest (T2). The data was analyzed using SPSS Statistics 24.

Descriptive Statistics: Raters Descriptive statistics were calculated for the sample of raters in terms of demographic variables, such as gender, age and art-related qualification.

Descriptive Statistics: Items Descriptive statistics were calculated for each item.

Item Difficulty Item difficulty reflects the difficulty to answer an item. As a measure for item difficulty Dahl's difficulty index for interval-levelled data was computed. It is the arithmetic mean of item responses and is defined within the range of 0 to 1. The higher its

value, the easier it is for raters to answer symptomatically, meaning to give a positive response. Vice versa, the lower the value, the less likely agreement is. While items with extreme indices can barely differentiate, since extreme characteristics are rarely given, those with a medium difficulty are the most informative (Moosbrugger & Kelava, 2007, p. 74).

Capacity of Differentiation Another criterion is that items should be able to distinguish between different types of images. For example, the difference between a tender pencil drawing and a pasty oil-painting should be reflected in the data. Capacity of differentiation between images was examined for each item, as well as the overall test by criteria like item variance and item effect size. Item variance is understood as an item's differentiation ability with regard to the sample, in this case the image sample (Moosbrugger & Kelava, 2007, p. 71). Item variance was calculated along with the descriptive statistics. Item effect size is a further measure for indicating the amount of variance explained by an item or the overall test. It was calculated using a multivariate analysis of variance (MANOVA), in which the items served as dependent variables and the images as fixed factors. The resulting partial eta-squared computed (Levine & Hullett, 2002) was interpreted as an effect size estimator for the variance explained by each item and the overall test.

Test-Retest Reliability One and the same picture should be rated similarly at any given time. As an indicator for stability and ratings not being arbitrary, test-retest reliability between both times of measurement was determined by calculating Pearson product-moment correlations.

Principal Component Analysis Factor analysis serves as a tool to examine whether items correlate highly with the factors they are supposed to measure, such as constructs, dimensions, or attributes. An exploratory procedure was chosen, since there is no previous empirical data available on a possible factors' structure. PCA was chosen as a method of extraction, because it aims at explaining as much variance of the observed variable as possible (Moosbrugger & Kelava, 2007, p. 308). A PCA followed by varimax rotation was conducted. This was based on recommendations by Brown (2009) and Suhr (2005) as an exploratory data reduction procedure (Osborne & Costello, 2004; Suhr, 2006). Kaiser criterion and scree plot were applied and results were compared to Horn's parallel analysis, using O'Connor's syntax (O'Connor, 2000, 2018). Factors with larger eigenvalues than parallel components were extracted in another PCA followed by varimax rotation.

Tucker's Coefficient of Congruence Tucker's coefficient of congruence (Lorenzo-Seva & Ten Berge, 2006; Tucker & Lewis, 1973) was computed, functioning as a further index on accepting or rejecting the given factor's solution. The coefficient represents the cosine of the angle between two vectors and a standardized measure of proportionality of

elements (Lorenzo-Seva & Ten Berge, 2006). It is an additional reliability coefficient indicating whether an otherwise acceptable factor model, such as the exploratory one above retrieved by a PCA, does indeed represent the interrelations among the attributes for a population (Tucker & Lewis, 1973).

The procedure was followed and the coefficient was calculated as outlined by Wünsch (2016) as well as Fetz, Vogt, Ostermann, Schmitz, and Schulz-Quach (2018). The dataset was split into two random samples, a PCA followed by varimax rotation was calculated, and the suggested eight factors were extracted. The resulting component loading matrices were both compared with each other. Two-tailed Pearson correlations ($\alpha = .05$) were calculated. Based on significant correlations between factors from both samples, factors were accordingly paired and Tucker's coefficient of congruency was calculated. The values of this statistical parameter per definition range between +1 and -1 and can be interpreted similarly to the Pearson's product-moment correlation (Chan, Ho, Leung, Chan, & Yung, 1999). Since it depicts the degree of deviation between two factors, negative coefficients can be interpreted similarly to positive values. Results were interpreted as recommended by Chan et al. (1999) and Lorenzo-Seva and Ten Berge (2006), where a critical congruence level, rather than unity for indication of identic factors is suggested. Values between .85 - .94 indicate fair similarity, whereas values higher than .95 suggest two factors or components being equal.

Inter-Rater Reliability Inter-rater agreement serves as a measure for objectivity. It was calculated for each item and the overall test as a measure for relative consistency between the expert ratings (LeBreton & Senter, 2008). On the basis of not all images being rated by each subject, and values being averaged, intraclass correlation coefficients (*ICCs*) one-way random average measures with absolute agreement were calculated. Due to the restructuring of the dataset by averaging beyond raters for the analysis of *ICCs*, missing values occurred that fulfill the assumptions of missing at random (MAR; Rubin, 1996; Schafer & Olsen, 1998). A frequency analysis of missing values per case was conducted. In consequence, three different models of datasets were compiled and the *ICCs* calculated separately with these were compared to each other. For reference, model A was created by excluding all cases with more than 20% missing values from the dataset. No missing values were imputed resulting in SPSS's basic procedure in treating missing values, listwise deletion. Since this diminished sample size resulted in a decrease of power of the analysis, two further models were compiled both using the Expectation-Maximization (EM) algorithm (Baltes-Götz, 2013; Pigott, 2001) for imputation. In model B, all cases with more than 20% missing values were excluded as well, and the remaining missing values were imputed. In model C, all cases were included and all

missing values were imputed. Multivariate normal data is assumed. EM algorithm was applied, because of its advantages in comparison to other simpler ways of treating missing values (e.g., listwise deletion or mean imputation), and for preserving the relationship with other variables (Pigott, 2001). Since there are no significant differences between Monte Carlo Markov chain (MCMC) method and EM algorithm in accuracy of imputation (Lin, 2010), we went for the latter, as the plainer procedure resulting in only one imputed dataset. Finally in addition Cronbach's alpha was calculated as a measure for internal consistency, which in this case is functionally equivalent to the formulation of ICCs (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

RESULTS

Descriptive Statistics: Raters

In total, 1,160 participants started the online survey at T1. 643 of those processed only three of the six images. However, we only used the data of raters who completed T1 and processed all six images they were presented with. T1 was completed by a total of 521 raters. T2 was completed by 239 of those. At T1, data of eleven raters were excluded from the analysis for admitting to not having filled out the survey seriously. Data of four more raters was excluded, because they did not meet the inclusion criteria of expertise. The final sample of valid raters at T1 consists of 506 experts (432 female, 66 male, 8 diverse) between 19 and 77 years ($M = 32.92$, $SD = 11.68$). All of them have an academic degree in an art-related subject (63.6 %) or are in at least in their second year of study (36.4 %). The raters' disciplines include art therapy (24.9%), art history (19.2 %), visual art (14.6 %), art pedagogy (12.8 %), graphic design (4.7 %), design (4.7 %), art sciences (3.4%), restoration (2.4 %), image sciences (.4 %), and various others (12.8 %). Each picture is rated by one to 19 raters ($M = 9.54$, $SD = 2.50$) at T1 and by zero to ten raters ($M = 4.49$, $SD = 1.89$) at T2.

Descriptive Statistics: Items

The calculated item means range from 1.81 to 3.29 ($M = 2.35$, $SD = .40$) at T1 and from 1.75 to 3.25 ($M = 2.34$, $SD = .38$) at T2. Item standard deviations lie between 1.20 and 1.75 ($M = 1.49$, $SD = .11$) at T1 and between 1.14 and 1.69 ($M = 1.44$, $SD = .12$) at T2 (see Table 2).

Item Difficulty

Item difficulty p_i ranges from .36 to .66 ($M = .47$, $SD = .08$) at T1 and from .35 to .65 ($M = .46$, $SD = .08$) at T2 (see Table 2).

Capacity of Differentiation

Capacity of differentiation in the form of items' variance s_i^2 ranges from 1.48 to 3.05 ($M = 2.24$, $SD = .33$) at T1 and from 1.31 to 2.84 ($M = 2.09$, $SD = .34$) at T2 (see Table 2). Par-

tial eta-squared ranges from .23 to .58 ($M = .45$, $SD = .09$) at T1 and from .35 to .71 ($M = .54$, $SD = .08$) at T2 (see Table 2). The overall test yields a partial eta-squared of .31 (Pillai's trace: $F(8243, 70642) = 3.80$, $p = .000$) at T1 and .40 (Pillai's trace: $F(8164, 28938) = 2.38$, $p = .000$) at T2.

Table 2

RizbA: Translated* items and item values (T1+T2), unit of analysis pictures

No. Item*			<i>M</i>	<i>SD</i>	<i>s_i²</i>	<i>p_i</i>	<i>η_p²</i>
1	The picture includes graphic elements	<i>T1</i>	2.91	1.53	2.33	.36	.38
		<i>T2</i>	2.87	1.48	2.20	.35	.50
2	The picture includes pictorial elements	<i>T1</i>	3.29	1.48	2.19	.36	.50
		<i>T2</i>	3.25	1.46	2.12	.36	.56
3	The manner of representation is concrete	<i>T1</i>	2.87	1.63	2.65	.36	.58
		<i>T2</i>	2.81	1.57	2.45	.36	.71
4	The manner of representation is abstract	<i>T1</i>	2.37	1.61	2.60	.37	.57
		<i>T2</i>	2.41	1.57	2.47	.37	.66
5	The color application is pastose	<i>T1</i>	1.82	1.43	2.04	.38	.38
		<i>T2</i>	1.98	1.44	2.06	.38	.47
6	The predominant coloring is vibrant	<i>T1</i>	2.24	1.52	2.31	.38	.53
		<i>T2</i>	2.29	1.46	2.14	.38	.63
7	In the picture primary colors are prevalent	<i>T1</i>	1.91	1.36	1.86	.39	.29
		<i>T2</i>	1.92	1.30	1.68	.38	.35
8	In the picture mixed colors (secondary colors) are prevalent	<i>T1</i>	2.69	1.49	2.22	.41	.46
		<i>T2</i>	2.70	1.45	2.11	.39	.54
9	In the picture there are complementary contrasts	<i>T1</i>	2.42	1.60	2.55	.41	.41
		<i>T2</i>	2.43	1.59	2.53	.42	.50
10	In the picture organic shapes are prevalent	<i>T1</i>	2.70	1.42	2.03	.44	.49
		<i>T2</i>	2.72	1.38	1.91	.43	.58
11	In the picture geometric shapes are prevalent	<i>T1</i>	2.05	1.48	2.20	.45	.56
		<i>T2</i>	1.95	1.40	1.97	.43	.64
12	The layout of the line is predominantly curved	<i>T1</i>	2.47	1.44	2.08	.45	.50
		<i>T2</i>	2.57	1.37	1.89	.45	.60
13	The layout of the line is predominantly angled	<i>T1</i>	1.83	1.45	2.11	.47	.54
		<i>T2</i>	1.81	1.38	1.91	.45	.63
14	The picture includes unworked areas	<i>T1</i>	1.92	1.75	3.05	.47	.53
		<i>T2</i>	1.93	1.69	2.84	.46	.59
15	The picture appears to be deep	<i>T1</i>	2.37	1.39	1.92	.48	.37
		<i>T2</i>	2.28	1.35	1.83	.48	.49
16	The picture is perspectival	<i>T1</i>	2.52	1.50	2.23	.50	.48
		<i>T2</i>	2.43	1.46	2.12	.48	.58
17	The picture is without perspective (aperspectival)	<i>T1</i>	2.08	1.60	2.56	.50	.44
		<i>T2</i>	2.19	1.55	2.41	.48	.58

No. Item*		<i>M</i>	<i>SD</i>	s_i^2	p_i	η_p^2
18 The picture is restless	<i>T1</i>	2.54	1.43	2.04	.51	.42
	<i>T2</i>	2.47	1.36	1.84	.50	.52
19 The picture is wild	<i>T1</i>	2.23	1.46	2.13	.51	.49
	<i>T2</i>	2.24	1.38	1.91	.51	.55
20 The global composition is laid out vertically	<i>T1</i>	2.49	1.57	2.45	.54	.48
	<i>T2</i>	2.40	1.53	2.35	.53	.57
21 The global composition is laid out horizontally	<i>T1</i>	1.96	1.52	2.32	.54	.45
	<i>T2</i>	1.92	1.47	2.16	.54	.56
22 The global composition is laid out diagonally	<i>T1</i>	1.82	1.48	2.20	.55	.40
	<i>T2</i>	1.75	1.42	2.03	.55	.49
23 The global composition is laid out area-wide without a main subject (All-Over-Structure)	<i>T1</i>	1.81	1.68	2.81	.56	.47
	<i>T2</i>	1.87	1.64	2.69	.55	.56
24 The picture appears to be diffuse	<i>T1</i>	2.20	1.41	2.00	.58	.31
	<i>T2</i>	2.20	1.34	1.80	.57	.43
25 The picture appears to be precise, accurate	<i>T1</i>	2.74	1.38	1.90	.58	.41
	<i>T2</i>	2.74	1.32	1.74	.57	.51
26 The picture appears to be harmonic	<i>T1</i>	2.78	1.22	1.48	.66	.23
	<i>T2</i>	2.69	1.14	1.31	.65	.36
<i>M</i>	<i>T1</i>	2.35	1.49	2.24	.47	.45
	<i>T2</i>	2.34	1.44	2.09	.46	.54
<i>SD</i>	<i>T1</i>	.40	.11	.33	.08	.09
	<i>T2</i>	1.75	1.14	1.31	.35	.35
<i>Min</i>	<i>T1</i>	1.81	1.22	1.48	.36	.23
	<i>T2</i>	1.75	1.14	1.31	.35	.35
<i>Max</i>	<i>T1</i>	3.29	1.75	3.05	.66	.58
	<i>T2</i>	3.25	1.69	2.84	.65	.71

* = original German version see additional material, T1 = test, T2 = retest, M = mean (0 = strongly disagree, 5 = strongly agree), SD = item standard deviation, Min = minimum, Max = maximum, s_i^2 = item variance, p_i = item difficulty, η_p^2 = partial eta-squared effect size estimator

Test-Retest Reliability

Overall test-retest reliability yields a Pearson product-moment correlation of .86 ($r(3510) = .86, p = .000$).

Principal Component Analysis

A PCA followed by varimax rotation for both T1 and T2 results in eight factors with eigenvalues above one referring to Kaiser criterion. Both scree plots suggest six factors. Compared to Horn's parallel analysis, eight factors have larger eigenvalues than parallel components at both times of measurement. In another PCA followed by varimax rotation, eight factors are extracted accounting for 63.65 % (T1) and 65.99 % (T2) of the common variance (see Table 3 and 4).

Table 3 RizbA: Rotated component matrix, factor loadings (T1)

Preliminary labels	No.	Item	Components							
			1	2	3	4	5	6	7	8
<i>shaping</i>	13	The layout of the line is predominantly angled	-.800							
	11	In the picture geometric shapes are prevalent	-.782							
	10	In the picture organic shapes are prevalent	.748							
	12	The layout of the line is predominantly curved	.725							
<i>picture effect</i>	18	The picture is restless		.834						
	19	The picture is wild		.784						
	26	The picture appears to be harmonic		-.695						
	24	The picture appears to be diffuse		.689						
	25	The picture appears to be precise, accurate		-.544						
<i>spatiality</i>	16	The picture is perspectival			.840					
	15	The picture appears to be deep			.830					
	17	The picture is without perspective (aperspectival)			-.743	.400				
<i>representation</i>	4	The manner of representation is abstract				.697				
	3	The manner of representation is concrete				-.672				
	23	The global composition is laid out area-wide without a main subject (All-Over-Structure)				.640				
<i>color intensity</i>	6	The predominant coloring is vibrant					.789			
	9	In the picture there are complementary contrasts					.702			
<i>pictorial elements (painting)</i>	8	In the picture mixed colors (secondary colors) are prevalent						.823		
	7	In the picture primary colors are prevalent					.526	-.663		
	2	The picture includes pictorial elements	.410					.441		
	5	The color application is pastose						.418		
<i>composition</i>	20	The global composition is laid out vertically							.875	
	21	The global composition is laid out horizontally							-.705	
<i>pictorial elements (drawing)</i>	1	The picture includes graphic elements								.688
	14	The picture includes unworked areas								.615
	22	The global composition is laid out diagonally								.530

Method of extraction: PCA; Method of rotation: Varimax, Kaiser normalization; Values < .40 suppressed

Table 4 *RizbA*: Rotated component matrix, factor loadings (T2)

Preliminary labels	No.	Item	Components							
			1	2	3	4	5	6	7	8
<i>picture effect</i>	18	The picture is restless	.856							
	19	The picture is wild	.794							
	24	The picture appears to be diffuse	.762							
	26	The picture appears to be harmonic	-.694							
	25	The picture appears to be precise, accurate	-.569							
<i>shaping</i>	13	The layout of the line is predominantly angled		-.815						
	12	The layout of the line is predominantly curved		.8						
	10	In the picture organic shapes are prevalent		.785						
	11	In the picture geometric shapes are prevalent		-.769						
<i>representation</i>	4	The manner of representation is abstract			.741					
	3	The manner of representation is concrete			-.740					
	23	The global composition is laid out area-wide without a main subject (All-Over-Structure)			.713					
<i>pictorial elements (painting)</i>	8	In the picture mixed colors (secondary colors) are prevalent				.726				
	2	The picture includes pictorial elements				.651				
	5	The color application is pastose				.585				
<i>pictorial elements (drawing)</i>	14	The picture includes unworked areas				-.577				
	1	The picture includes graphic elements				-.491	.456			
<i>spatiality</i>	15	The picture appears to be deep					.786			
	16	The picture is perspectival			-.467		.747			
	17	The picture is without perspective (aperspectival)			.536		-.634			
<i>color intensity</i>	6	The predominant coloring is vibrant						.760		
	7	In the picture primary colors are prevalent				-.422		.723		
	9	In the picture there are complementary contrasts						.634		
<i>composition</i>	21	The global composition is laid out horizontally							-.820	
	20	The global composition is laid out vertically							.790	
	22	The global composition is laid out diagonally								.913

Method of extraction: PCA; Method of rotation: Varimax, Kaiser normalization; Values < .40 suppressed

Tucker's Coefficient of Congruence

Tucker's coefficient of congruence ranges between $|.97|$ and $|1.00|$ at T1 and between $|.71|$ and $|.99|$ at T2 (see Table 5).

Table 5
Tucker's coefficient of congruence (T1 and T2)

Time of measurement	Factor	Tucker's coefficient of congruence
T1	factor ₁₁	.99
	factor ₂₂	1.00
	factor ₃₄	.99
	factor ₄₃	.98
	factor ₅₅	.99
	factor ₆₆	.97
	factor ₇₇	.99
	factor ₈₈	.98
T2	factor ₁₂	-.92
	factor ₂₃	.99
	factor ₃₁	.98
	factor ₄₄	.96
	factor ₅₆	.83
	factor ₆₇	.97
	factor ₇₈	.94
	factor ₈₅	.71

Inter-Rater Reliability

For model A items' ICCs range from .55 to .91 ($M = .82$, $SD = .08$) at T1 and from .32 to .82 ($M = .66$, $SD = .11$) at T2. Overall ICCs are .86 ($F(6681, 46774) = 7.14$, $p = .000$, 95% CI [.85, .87]) at T1 and .71 ($F(7045, 14092) = 3.50$, $p = .000$, 95% CI [.70, .73]) at T2. For model B items' ICCs range from .57 to .90 ($M = .83$, $SD = .07$) at T1 and from .36 to .82 ($M = .68$, $SD = .10$) at T2. Overall ICCs are .86 ($F(8267, 57876) = 7.38$, $p = .000$, 95% CI [.86, .87]) at T1 and .73 ($F(8267, 16536) = 3.67$, $p = .000$, 95% CI [.72, .74]) at T2. For model C items' ICCs range from .87 to .97 ($M = .94$, $SD = .02$) at T1 and from .73 to .92 ($M = .87$, $SD = .04$) at T2. Overall ICCs are .96 ($F(8267, 148824) = 22.94$, $p = .000$, 95% CI [.96, .96]) at T1 and .90 ($F(8267, 82680) = 9.49$, $p = .000$, 95% CI [.89, .89]) at T2 (see Table 6; Figure 2 and 3).

Table 6
RizbA: Translated* items and item values (T1+T2), unit of analysis raters

No.	Item*		<i>model A</i>	<i>model B</i>	<i>model C</i>
1	The picture includes graphic elements	<i>T1</i>	.77	.78	.93
		<i>T2</i>	.65	.67	.84
2	The picture includes pictorial elements	<i>T1</i>	.86	.87	.96
		<i>T2</i>	.70	.72	.84
3	The manner of representation is concrete	<i>T1</i>	.91	.90	.97
		<i>T2</i>	.82	.82	.91
4	The manner of representation is abstract	<i>T1</i>	.90	.89	.97
		<i>T2</i>	.76	.77	.92
5	The color application is pastose	<i>T1</i>	.79	.80	.93
		<i>T2</i>	.57	.62	.87
6	The predominant coloring is vibrant	<i>T1</i>	.88	.88	.96
		<i>T2</i>	.76	.77	.91
7	In the picture primary colors are prevalent	<i>T1</i>	.69	.72	.90
		<i>T2</i>	.32	.36	.80
8	In the picture mixed colors (secondary colors) are prevalent	<i>T1</i>	.84	.84	.95
		<i>T2</i>	.67	.67	.86
9	In the picture there are complementary contrasts	<i>T1</i>	.80	.82	.94
		<i>T2</i>	.60	.62	.88
10	In the picture organic shapes are prevalent	<i>T1</i>	.86	.86	.96
		<i>T2</i>	.73	.73	.86
11	In the picture geometric shapes are prevalent	<i>T1</i>	.90	.90	.97
		<i>T2</i>	.77	.77	.92
12	The layout of the line is predominantly curved	<i>T1</i>	.87	.87	.96
		<i>T2</i>	.73	.73	.88
13	The layout of the line is predominantly angled	<i>T1</i>	.89	.89	.96
		<i>T2</i>	.75	.76	.92
14	The picture includes unworked areas	<i>T1</i>	.89	.89	.96
		<i>T2</i>	.73	.75	.92
15	The picture appears to be deep	<i>T1</i>	.77	.78	.93
		<i>T2</i>	.60	.63	.86
16	The picture is perspectival	<i>T1</i>	.85	.85	.95
		<i>T2</i>	.69	.71	.89
17	The picture is without perspective (aperspectival)	<i>T1</i>	.83	.84	.95
		<i>T2</i>	.66	.69	.90
18	The picture is restless	<i>T1</i>	.81	.82	.94
		<i>T2</i>	.66	.68	.86
19	The picture is wild	<i>T1</i>	.85	.86	.96
		<i>T2</i>	.71	.72	.88

No.	Item*		model A	model B	model C
20	The global composition is laid out vertically	T1	.86	.86	.96
		T2	.71	.73	.89
21	The global composition is laid out horizontally	T1	.82	.85	.95
		T2	.67	.69	.91
22	The global composition is laid out diagonally	T1	.82	.81	.94
		T2	.60	.63	.89
23	The global composition is laid out area-wide without a main subject (All-Over-Structure)	T1	.85	.85	.95
		T2	.69	.69	.91
24	The picture appears to be diffuse	T1	.70	.72	.91
		T2	.62	.61	.84
25	The picture appears to be precise, accurate	T1	.81	.81	.94
		T2	.64	.68	.84
26	The picture appears to be harmonic	T1	.55	.57	.87
		T2	.42	.47	.73
<i>M</i>		T1	.82	.83	.94
		T2	.66	.68	.87
<i>SD</i>		T1	.08	.07	.02
		T2	.11	.10	.04
<i>Min</i>		T1	.55	.57	.87
		T2	.32	.36	.73
<i>Max</i>		T1	.91	.90	.97
		T2	.82	.82	.92

* = original German version see additional material, T1 = test, T2 = retest, *M* = mean (0 = strongly disagree, 5 = strongly agree), *SD* = item standard deviation, *Min* = minimum, *Max* = maximum, *ICC* = intraclass correlation coefficient average measure

Figure 2. Items' ICCs (T1): Model A, B, and C with 95% confidence interval

Figure 3. Items' ICCs (T2): Model A, B, and C with 95% confidence interval

DISCUSSION

In empirical art psychology and creativity research most studies focus on the psychological correlates of art. Only a few go beyond treating artworks as categorical data (e.g. abstract vs. representational) and consider artworks in detail. The rating instrument for two-dimensional pictorial works *RizbA* makes a difference to current research approaches. This article outlines the validity of the novel *RizbA* by means of a statistical approach. Thus, the results presented here need to be discussed by both statistical quality criteria, as well as their limitations and implications given for further research.

Statistical Quality Criteria

Item Difficulty The items yield medium-sized difficulty indices, providing excellent information on and differentiation between pictorial characteristics.

Capacity of Differentiation Item variances as an additional criterion for capacity of differentiation are moderate to large. Comparing partial eta-squared to common effect sizes (Coster, 2012), large effects can be assumed here as well, meaning the instrument is capable of differentiating between different types of images.

Test-Retest Reliability Relating to common effect size conventions (Cohen, 1992), the overall test-reliability suggests a large effect and implies *RizbA* as being a stable measure across time. This ensures ratings are not arbitrary, but one and the same picture is rated similar at any given time.

Principal Component Analysis

The results empirically support the existence of an interpretable factor structure. PCA suggests eight interpretable components, measured by the scale, that account for a large part of common variance. Prospective factor labels were *a posteriori* designated based on the empirical data and theoretical reasoning. This involved an iterative discussion process of expert opinions referring to art theory, which included experts of the HKS – University of Applied Sciences and Arts in Ottersberg and the German Creative Arts Therapies Research Association (FVKT).

Resulting factor labels (in corresponding order to Table 5) might be *shaping*, *picture effect*, *spatiality*, *representation*, *color intensity*, *pictorial elements (painting)*, *composition*, and *pictorial elements (drawing)*. The number of extracted factors is different from the first pilot study, which indicated only four factors preliminarily interpreted as *picture effect*, *representation*, *shaping* and *perspective* (Schoch et al., 2017). However, they resemble the factors structure in the second study that used amateur pictorial works (Schoch & Ostermann, under review), with *shaping* and *picture effect* factors being the most stable components across studies. The *picture effect* factor is stable across times of measurement

and studies and contains high factor loadings. However, in terms of content validity, it does not clearly measure the defined concept of formal picture elements. Thought must be given to exclude this factor in the final test version. Moreover, especially two items should be further reviewed: Item 7 (*In the picture primary colors are prevalent*) on both times of measurements has diametrically opposed cross loadings on two factors, namely *color intensity* and *pictorial elements (painting)*. A theoretical reason to explain this cross-load might be that color is a relevant item for both factors. Item 22 (*The global composition is laid out diagonally*) has a medium loading on *pictorial elements (drawing)* at T1, but turns into a highly loading single item factor at T2.

The most intriguing conclusion of the expert group process is that the empirical factor solution does not properly reflect existing art theories as described in professional literature (e.g., Belting, Dilly, Kemp, Sauerländer, & Warnke, 1996). This becomes particularly clear when comparing the factor labels of this study with the theoretical content areas used to compile the initial pool of 113 items in early test construction phase: *representation, color, shape, space, movement, composition* and *expression* (Schoch, 2014). Those might sound similar to current labels, but contain very different items than a priori theoretically hypothesized. One explanation for this disparity might be that *RizbA* does not measure formal elements in terms of a theoretical classification model of picture analysis, like art history, or image science contextualizes them. Rather, *RizbA* appears to capture sort of a temper or atmosphere of the image itself. However, it is not a subjective impression of how an image is perceived. It is intersubjectively measurable as the inter-rater reliabilities indicate. This might suggest a gap between art theory and its empirical findings that should be further addressed.

Tucker's Coefficients of Congruence Tucker's coefficients of congruence at T1 all indicate high similarity and thus all respective two components being equal. At T2, two factors yield high, four yield fair, and two suggest only moderate similarity. Still, the results obtained are rather conservative (Wünsch, 2016). Analogically to the previous study, we went for the conservative approach in contrast to the usual procedure, stipulating a Procrustean factor rotation with orthogonal rotation to maximum similarity (Chan et al., 1999; Cliff, 1966; McCrae, Zonderman, Costa Jr, Bond, & Paunonen, 1996; Raykov & Little, 1999; Schönemann, 1966; Zumbo, Sireci, & Hambleton, 2003).

Inter-Rater Reliability Comparing ICCs resulting from the three datasets, model C yields the highest inter-rater reliability. Still, due the amount of values imputed, the dataset includes zero variance items and is likely to overestimate the effect. Reliabilities in model A and B are comparable with one another, but generally lower than in model C. Model A

suffers from the disadvantages of listwise deletion, such as biased estimations and loss of precision (Baltes-Götz, 2013). However, to avoid these and prevent a decrease in power we decided to go for model B, which entails a moderate loss of data due to excluding cases, but provides a realistic dataset with sufficient power. Based on statistical guidelines (Koo & Li, 2016; LeBreton & Senter, 2008), *ICC* of the overall test at T1 can interpreted as strong agreement among raters (LeBreton & Senter, 2008) and thus good inter-rater reliability (Koo & Li, 2016). On the item level, the results reach from moderate to very strong agreement and hence moderate to good reliability. At T2, the entire test yields strong agreement and moderate reliability. On the item level, results are more diverse reaching from weak to strong agreement, which suggests poor to good reliability.

When compared, the test quality at T1 is mostly higher and more reliable than at T2, which may be because of the differences in the sample size of raters. This is reflected in good to excellent results at T1 compared to considerably weak to moderate quality criteria at T2 regarding *ICCs*, Tucker's coefficients of congruence and PCA in particular (Suhr, 2005). Considering the partly small amount of ratings for some stimuli at T2, it is debatable to what extent results of the retest are conclusive for other quality criteria than test-retest reliability.

Comparing the present results to those of the previous validation study by Schoch and Ostermann (under review) on amateur art, item means, item standard deviations, item variance, and partial eta-squared reach similar values. Factorial structures resemble one another and Tucker's coefficient of congruence are comparably high. *ICCs* for the overall test are similar in both studies, with T1 of the current study yielding a slightly higher reliability compared to T1 of the previous study. Then again, T2 of the current study has slightly less reliability than at T2 in the previous study. The only clear difference lies within test-retest reliability, which was .93 higher in the preceding study compared to the current one, attaining only .86.

Limitations

Technical Aspects There are two technical drawbacks as result of the online data collection. Image sizes were rather small to meet the *SoSci Survey's* (Leiner, 2018) maximum upload capacity and to make sure images load fast with any internet connection. Based on the large sample size needed, this was discarded. Also, the raters viewed the images on their own devices, which might result in a different display of color and resolution. Using a better quality of image solution, a different rating experience may be expected, especially when original works instead of digital copies are used.

As already mentioned, for some images at the retest, the number of ratings is insufficiently small. In the survey, images were drawn without putting back, which should normally lead to a reasonably equal number of ratings per picture. However, if raters did not terminate the survey, the images presented to them were already drawn, but are not represented in the data, because only complete sets were included in the analysis. Future studies should use an even more detailed PHP code, putting pictures back if subjects prematurely abort the survey. However, the general sample size at T2 was not sufficiently large to calculate all statistics appropriately with sufficient power. Results at T2 are to be interpreted with caution.

Data Structure

As described, datasets regarding raters' agreement were merged, which resulted in missing values. For not losing too much data through listwise deletion, in model B only those generic raters with less than 20 % missing values were used and the remaining missing values were imputed. Since the presentation of artworks was randomized and the assumption of missing at random is fulfilled, the results are probably not biased by this *a posteriori* defined cut-off point. Still, the criterion was not defined *a priori* and there is a non-negligible loss of information. Also, there are several disadvantages of imputation compared to originally obtained data, for example underestimating the standard error (Baltes-Götz, 2013).

Factor Analysis

PCA suggests an interpretable factor structure with eight components. Still, not having conducted a confirmatory factor analysis (CFA), final conclusions cannot be drawn. On the basis of the exploratory state of research in this field, other statistical criteria usually used for test validation were not yet applied. This includes internal consistency measures, such as split-half reliability, which require test scales measuring one construct, to date an uncertain condition.

Samples

The metadata about the images provided by *prometheus* varies slightly in terms of completeness, precision, and systematics, due to stemming from different institutions, which contribute to the archive. Although these sources are institutes, research facilities and museums, it cannot be guaranteed that all information and dating is unflawed. Due to an error, two artists' images were included in the sample, opposed to the initial regulation specifying one image each.

Although the amount of information included in the metadata differs, it reveals several biases regarding ethnicity and gender in image databases. In principle, the pictorial material includes a relatively heterogeneous sample created by artists of different origins.

The *prometheus* database includes works by international artists from European, African, American, and Asian countries. Still, being based in Germany, there is a bias towards a western cultural area. Considering the rater sample, they also come from a probably Eurocentric cultural context. Since art and several psychological variables differ across cultures (Cattaneo, 1994), it can be assumed that there are cultural differences in art-making and art perception as well. Speaking about the artists' gender, 73.4 % of its creators are male, only 18.4 % female and 0.3 % diverse. For the remaining 7.9 % the gender is not assignable based on the metadata. Unfortunately, this bias in gender and diversity reflect the current art world (Fraiberger, Sinatra, Resch, Riedl, & Barabási, 2018), where cis white males are still disproportionally more featured. Network effects play a crucial role in an artist's reputation and valuation, while being from a country with better access to the art network means having higher chances of starting and ending one's career at the top. (Fraiberger et al., 2018). Considering the rater sample, experts stem from German speaking countries (Germany, Austria, Switzerland). Here, 85.38 % are female, 13.04 % male and 1.58 % diverse.

Thus, the image and the rater sample are both biased in terms of gender and tend towards a western, in particular Eurocentric, and probably objectivistic cultural context. Since art as well as several psychological variables differ across cultures (Cattaneo, 1994), it is highly likely to find differences in art-making and rating art as well. This probably also refers to gender and other contexts relevant to both pictures and raters. Future studies should investigate the instrument's validity by taking into account gender, ethnicity, and other relevant variables of both pictures and raters with a critical view on Eurocentrism (Mosquera, 1992) and paternalistic structures. In particular, scientific databases and research should strive for more diversity to accurately represent the spectrum of the population. Although at the time of this study *WikiArt* (<https://www.wikiart.org>) is not comprehensive enough when it comes to contemporary art, open projects like this with specific categories such as "female artists" might become a promising tool in conducting art-related research (e.g., Marengo, Fazekas, & Tombros, 2017; Mohammad & Kiritchenko, 2018).

Artworks

In the current study, only two-dimensional artworks were used as image material. It is apparent that contemporary visual art entails a multitude of art forms, such as plastics, installations, environments, happenings, digital art, street art, and more. Addressing this issue there has currently been an item pool compiled for three-dimensional works (e.g., sculptures, plastics, reliefs). It is waiting to be empirically tested. However, more specific instruments are needed to measure walk-in artworks in terms of a formal analysis analogously to *RizbA*.

Implications

RizbA was successfully validated on a sample of contemporary art. Taking the study's results into consideration, it is highly likely that they can be generalized to the entirety of pictorial contemporary art. This makes *RizbA* a valid and reliable instrument for all sorts of handmade two-dimensional pictorial works – whether created by artists or amateurs.

Enhancing Methodology in Research on Art

Independently from the focus of research on art, it would be highly desirable not only to capture psychological and physiological variables, but also the artwork itself. Until now, there has been a lack of instruments that are capable of reliably describing these. According to our results, *RizbA* might make a difference. With *RizbA*, it may be possible to reliably examine both the artwork and its psychological correlates. As the first quantitative and reliable tool for formal picture analysis, *RizbA* allows a detailed analysis of the formal content and a wide range of applicability. In art psychology, objects of research could be further assessed, giving art-making and artworks significantly more weight. Adding such a measurement of artworks to existing methodology in the field, it might become highly instructive for a wide range of research topics, such as aesthetics, creativity, and personality.

Empirical Aesthetics

Psychological aesthetics has a long empirical tradition (Leder, Belke, Oeberst, & Augustin, 2004; Martindale, 2007) going back to experimental psychologists like Wundt (1874) and Fechner (1876). Still, the focus of empirical aesthetics is art perception with all its constructs, such as aesthetic experience, aesthetic appreciation, or aesthetic emotions. There are complex theoretical models like the *Information-processing Stage Model of Aesthetic Processing* (Leder et al., 2004), stating aesthetic experience to be a reciprocal cognitive and affective process, or the *Vienna Integrated Model of Art Perception* (Pelowski, Markey, Forster, Gerger, & Leder, 2017) – a model of processing visual artworks.

Outcomes provoked by an art object, such as aesthetic emotions (Menninghaus et al., 2018), aesthetic appreciation, and aesthetic judgement, have already been empirically examined. For example, people differ in art preference, their emotions towards it, and in what they regard to be art (Leder et al., 2012). Another example are field studies; for example, in museum contexts, where Tröndle and Tschacher (2012) found artworks being correlated with physical reactions of viewers, their spatial behavior, and their aesthetic ratings. Innovative field research like this could also be enriched by rating the art and including this data.

Existing theories and research of this sort could be supplemented with an additional view on the artworks' formal content. For example, an area for investigation could focus

on the part of a painting that makes people appreciate it – this might be the intensity of color or the dynamics – and not just which category – abstract or representational. One way to operationalize this could be by combining the *Art Reception Survey* (Hager et al., 2012) and *RizbA*, and assessing intercorrelations between them. For example, cognitive stimulation might be linked to an artwork's composition. Furthermore, in field studies one could additionally research whether there are specific contents of artworks that make people move differently in a gallery space.

For example, an artwork by Kara Walker with a lot of figures in movement and thus more dynamics might invite perceivers to behave differently than the quiet-color field painting by Helen Frankenthaler. This could be operationalized in a correlation study, in which both are rated, the artworks using *RizbA* and the viewers' movement profiles in the gallery space. And then again, a color field by Helen Frankenthaler might provoke a different response than a conceptual work by Tracy Emin. Hence, further instruments should be developed, each specifically capturing different types of art.

Aesthetics and Creativity

Still, research usually focuses on creativity and aesthetics separately, which limits the view on art inherently comprising both perspectives – the artist's and the art viewer's. On a theoretical level, Tinio (2013) developed the *Mirror Model of Art* to overcome this divide. The model connects art-making and art-viewing and proposes that aesthetic experiences mirror the art-making process. It proposes three levels in the art-making process (initialization / expansion and adaption / finalizing) reversely mirroring three stages in the process of art-experiencing (early automatic processing / intermediate memory-based processing / meaning-making, aesthetic judgements, aesthetic emotions).

In their research, Pelowski et al. (2017) investigate affective, evaluative, and neurophysiological correlates of perceiving and interacting with art. One could drive this idea that viewing and making art are connected and include sense-making processes (Sullivan & McCarthy, 2009), forwarded by designing a study that involves both creation and perception. Psychological measures and *RizbA* ratings could be used for investigating how inner processes and their pictorial expression in form of artworks are linked. Using *RizbA*, studies on those issues could go beyond measuring solely psychological variables and additionally capture the artworks themselves while they are being created and perceived.

Art and Personality

Although aesthetics, creativity, and art-making have to be differentiated, they correlate and overlap to a large extent (Zaidel, Nadal, Flexas, & Munar, 2013). For example, there is evidence for correlations between art preferences, thinking styles, and personality

(Gridley, 2006, 2013). Personality also predicts aesthetic chills, well before intelligence or demographic background (Silvia & Nusbaum, 2011).

Besides that, creativity is closely related with openness to experience. Creative people tend to have an affinity or ability for changing within their personality traits. Complexity might even be a trait inherent to creative personalities (Haller & Courvoisier, 2010). These findings relate to the paradoxes of creativity, one of which is that the combination of personality features do not necessarily or logically belong together; for example, divergent and convergent thinking (Cropley, 1997). This also relates to previous studies (Gridley, 2006, 2013), in which evidence was found for correlations between personality and art preference, for example when it comes to openness to experience.

Taking into account that aesthetics and creativity are linked and both are related to personality, it is likely that there are correlations between artworks as an outcome of these and personality as well. If openness to experience correlates positively with a liking for abstract art (Gridley, 2013), people with high scores might not only be more likely to appreciate, but also more likely to create abstract artworks. This could be assessed by methodologically combining *RizbA* with personality measures, such as HEXACO-PI-R (Lee & Ashton, 2009; Moshagen, Hilbig, & Zettler, 2014).

Analogies With Other Arts

Another direction of research would be to find similarities and differences between the arts, for example visual arts, performing arts, literature, and music. Research already conducted on music (Colver & El-Alayli, 2016), film (e.g., Hanich, Wagner, Shah, Jacobsen, & Menninghaus, 2014; Wassiliwizky, Wagner, Jacobsen, & Menninghaus, 2015), or poetry (e.g., Menninghaus, Bohrn, Altmann, Lubrich, & Jacobs, 2014) could be expanded to visual arts. For example, analogical to Wassiliwizky et al. (2015) investigating art-elicited chills with regard to films, one could examine which formal elements in an image move you and make you shiver – the corporeality in Louise Bourgeois' or rather the veiled colors in Esther Teichmann's works.

Future Research

The next obligatory step is to conduct CFA to verify the factor structure of the scale (Suhr, 2006). In terms of further investigating the previously mentioned gap between art theory and its empirical findings, the CFA should test two models against each other. The first model suggests a factor structure based on the empirical data collected. The other compounds items to factors, as art theory would predict dimensions. Finding substantial differences between theory and empirics would fuel the discourse between arts, humanities, and science described by Mersch (2019) and other theorists. Therefore, we are currently

conducting another study, using new image material and computing the CFA based on an evolved theoretical model of the construct pictorial expression.

Another consecutive step is translating and validating the questionnaire into English and making it accessible to a broader audience. Future research will also include other populations of raters, such as art experts from other cultural areas or raters from other professions than the arts. Addressing the latter issue, a study has currently been conducted, testing the questionnaire on expert vs. non-expert raters in art (Jerusalem, 2020). The resulting test manual explains art-related terminology and ensures that the questionnaire is also applicable by researchers and practitioners from other domains without art training (e.g., psychology, pedagogy, psychiatry). In terms of test validation, there will be further studies examining convergent and divergent validity by multitrait-multimethod matrix analyses (Campbell & Fiske, 1959) using related but distinct scales, such as the *Assessment of Art Attributes* (Chatterjee et al., 2010), the *Diagnostic Drawing Series* (B. M. Cohen & Mills, 2015; Mills, 2003) or the *Formal Elements Art Therapy Scale* (Gantt, 2016).

In order to achieve a maximum of comparability, statistical analyses were conducted analogically to the previous study, including averaging data to the relevant level of analysis. Future approaches might put more emphasis on the different units of analysis by using other statistical approaches, for example multilevel modelling (Silvia, 2007). Future research will concentrate on further validation regarding other sorts of two-dimensional pictorial works, such as non-handmade techniques, images created by children and adolescents and by people from non-western cultural contexts. The validation should also reach out to other populations of raters, for example art experts from other cultural areas or raters from professions other than the arts. Addressing the latter issue, a study is currently being conducted testing the questionnaire on expert vs. non-expert raters and developing a test manual explaining the terminology. This will make the questionnaire applicable to researchers and practitioners from all domains.

Due to the Open Data provided by RizbA validation studies, images and associated RizbA ratings can be further used for secondary data analyses. So far, there are similarities with a data set. However, RizbA is primarily a questionnaire, not a database. Firstly, because the image samples of contemporary art and amateur works collected so far are not necessarily representative for all types of artworks and not appropriate for all research questions. Secondly, RizbA's aim is to be a universal and open methodological tool that can be used on any sample of two-dimensional pictorial works on any research question it is appropriate for. Based on good to high values in inter-rater reliability in all previous validation studies, RizbA ratings can be considered stable and generalizable

among different rater samples. A further validation study using additional explanations of technical terms showed that even within raters without art expertise, good inter-rater agreement is given (Jerusalem, 2020). In practical terms, if we conducted two identical studies on the same image material with different rater samples, taken aside natural variation in data, the ratings would be alike, accurate and stable. Finally, *RizbA*'s focus does not rely on the raters' opinions, which are rather a method of gaining data. Images and their formal analysis by raters are the general unit of analysis here.

CONCLUSION

Although a final conclusion to the factors structure cannot be drawn yet, results are very promising. The questionnaire yields fair to high test quality regarding reliability and validity. After successfully validating *RizbA* on amateur and artist handmade two-dimensional pictorial works, its use can be recommended for both non-professional and professional contemporary art. Being a universal tool for quantitative formal picture analysis, it complements current research methodology in art and science. Future studies in art psychology, creativity research, and empirical aesthetics can go beyond focusing solely on psychological correlates. By the inclusion of a reliable measurement of picture analysis, art is given greater weight in art psychology. Future research will focus on a validation on other more diverse samples regarding image material and raters. A further goal is getting insight in the factors structure, perhaps revealing a Big Eight of visual art.

REFERENCES

- Baltes-Götz, B. (2013). Behandlung fehlender Werte in SPSS und Amos [Dealing with missing values in SPSS and Amos]. Universität Trier.
- Belting, H., Dilly, H., Kemp, W., Sauerländer, W., & Warnke, M. (1996). *Kunstgeschichte: Eine Einführung* [Art history: An introduction]. In H. Belting, H. Dilly, W. Kemp, W. Sauerländer, & M. Warnke (Eds.) (6th ed.). Reimer.
- Brown, J. D. (2009). Principal components analysis and exploratory factor analysis: Definitions, differences, and choices. *Statistics*, 13(1).
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81.
- Cattaneo, M. (1994). Addressing culture and values in the training of art therapists. *Art Therapy*, 11(3), 184-186. <https://doi.org/10.1080/07421656.1994.10759081>
- Chan, W., Ho, R. M., Leung, K., Chan, D. K.-S., & Yung, Y.-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods*, 4(4), 378.
- Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, 28(2), 207-222. <https://doi.org/10.2190/EM.28.2.f>

- Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, 31(1), 33-42.
- Cohen, B. M., & Mills, A. (2015). The Diagnostic Drawing Series (DDS) at thirty: Art therapy assessment and research. In D. Gussak & R. M Rosal, *The Wiley handbook of art therapy* (p. 558-568). Wiley-Blackwell.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155.
- Colver, M. C., & El-Alayli, A. (2016). Getting aesthetic chills from music: The connection between openness to experience and frisson. *Psychology of Music*, 44(3), 413-427. <https://doi.org/10.1177/0305735615572358>
- Commare, L., Rosenberg, R., & Leder, H. (2018). More than the sum of its parts: Perceiving complexity in painting. *Psychology of Aesthetics, Creativity, and the Arts*, 12(4), 380. <https://doi.org/10.1037/aca0000186>
- Coster, J. D. (2012). Converting effect sizes. <http://www.stat-help.com/spreadsheets.html>
- Cropley, A. (1997). Creativity: A bundle of paradoxes. *Gifted and Talented International*, 12(1), 8-14.
- Epstein, C. (2019). *Bildlicher Ausdruck von Depression: Erprobung des Ratinginstruments RizbA an Bildstichproben aus dem klinischen Kunsttherapiesetting [Pictorial expression of depression: Testing the ratinginstrument RizbA on image samples from a clinical art therapy setting]*. Witten/Herdecke University. Zenodo database. <https://doi.org/10.5281/zenodo.3365921>
- Fechner, G. T. (1876). *Vorschule der Ästhetik* [Preschool of aesthetics] (Vol. 1). Breitkopf & Härtel.
- Fetz, K., Vogt, H., Ostermann, T., Schmitz, A., & Schulz-Quach, C. (2018). Evaluation of the palliative symptom burden score (PSBS) in a specialised palliative care unit of a university medical centre-a longitudinal study. *BMC palliative care*, 17(1), 92. <https://doi.org/10.1186/s12904-018-0342-0>
- Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C., & Barabási, A.-L. (2018). Quantifying reputation and success in art. *Science*. <https://doi.org/10.1126/science.aau7224>
- Gantt, L. (2016). The Formal Elements Art Therapy Scale (FEATS). In D. Gussak & M. Rosal (Eds.), *The Wiley handbook of art therapy* (pp. 569-578). Wiley-Blackwell.
- GLAM Wiki. (2020). GLAM Outreach Wiki: Galleries, Libraries, Archives, Museums. <https://outreach.wikimedia.org/wiki/GLAM>
- Gridley, M. C. (2006). Concrete and abstract thinking styles and art preferences in a sample of serious art collectors. *Psychological Reports*, 98(3), 853-857.
- Gridley, M. C. (2013). Preference for abstract art according to thinking styles and personality. *North American Journal of Psychology*, 15(3), 463-481.

- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265-275.
- Hager, M., Hagemann, D., Danner, D., & Schankin, A. (2012). Assessing aesthetic appreciation of visual artworks: The construction of the Art Reception Survey (ARS). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4), 320-333. <https://doi.org/10.1037/a0028776>
- Haller, C. S., & Courvoisier, D. S. (2010). Personality and thinking style in different creative domains. *Psychology of Aesthetics, Creativity, and the Arts*, 4(3), 149-160. <https://doi.org/10.1037/a0017084>
- Hanich, J., Wagner, V., Shah, M., Jacobsen, T., & Menninghaus, W. (2014). Why we like to watch sad films. The pleasure of being moved in aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 8(2), 130-143. <https://doi.org/10.1037/a0035690>
- Hayn-Leichsenring, G. U., Kenett, Y. N., Schulz, K., & Chatterjee, A. (2020). Abstract art paintings, global image properties, and verbal descriptions: An empirical and computational investigation. *Acta Psychologica*, 202, 102936. <https://doi.org/10.1016/j.actpsy.2019.102936>
- Ishizu, T., & Zeki, S. (2011). Toward a brain-based theory of beauty. *PloS One*, 6(7), <https://doi.org/10.1371/journal.pone.0021852>.
- Janßen, B. (2018). *Erprobung des Ratinginstruments für zweidimensionale bildnerische Arbeiten (RizbA): Ansatz zu einer möglichen Untersuchung des bildnerischen Ausdrucks von Schmerz in gemalten Bildern von Menschen mit chronischer Schmerzsymptomatik [Testing the rating instrument for two-dimensional pictorial works (RizbA): Pilot study for investigating pictorial expression in pictures painted by people with chronic pain symptoms]*. HKS – University of Applied Sciences and Arts, Ottersberg. Zenodo database. <https://doi.org/10.5281/zenodo.3407808>
- Jerusalem, J. L. L. (2020, February 20). *Rizba für alle: Überprüfung der Einsatzmöglichkeiten eines Ratinginstruments für zweidimensionale bildnerische Arbeiten (RizbA) und des neu dazu erstellten Manuals [Rizba for everyone: Review of the possible uses of a rating instrument for two-dimensional pictorial works (RizbA) and the newly created manual]*. Witten/Herdecke University, Witten. Open Science Framework. <https://doi.org/10.17605/OSF.IO/FVJZM>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>

- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815-852. <https://doi.org/0.1177/1094428106296642>
- Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4), 489-508. <https://doi.org/10.1348/0007126042369811>
- Leder, H., Gerger, G., Dressler, S. G., & Schabmann, A. (2012). How art is appreciated. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1), 2-10. <https://doi.org/10.1037/a0026396>
- Leder, H., Tinio, P. P., Brieber, D., Kröner, T., Jacobsen, T., & Rosenberg, R. (2018). Symmetry Is Not a Universal Law of Beauty. *Empirical Studies of the Arts*, 37(1), 104-114. <https://doi.org/10.1177/0276237418777941>
- Lee, K., & Ashton, M. C. (2009). The HEXACO Personality Inventory - Revised: A measure of the six major dimensions of personality. <http://hexaco.org>
- Leiner, D. (2018). SoSci Survey. <https://www.soscisurvey.de>
- Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, 28(4), 612-625.
- Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality & Quantity*, 44(2), 277-287. <https://doi.org/10.1007/s11135-008-9196-5>
- Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57-64. <https://doi.org/10.1027/1614-1881.2.2.57>
- Maclagan, D. (1999). Getting the feel: Problems of research in the fields of psychological aesthetics and art therapy - The struggle with incarnation & the negative sublime. *The Arts in Psychotherapy*, 5(26), 303-311.
- Marengo, L., Fazekas, G., & Tombros, A. (2017). *The Interaction of Casual Users with Digital Collections of Visual Art. An Exploratory Study of the WikiArt Website*. International Conference on Human-Computer Interaction. Springer.
- Martindale, C. (2007). Recent trends in the psychological study of aesthetics, creativity, and the arts. *Empirical Studies of the Arts*, 25(2), 121-141. <https://doi.org/10.2190/B637-1041-2635-16NN>
- McCrae, R. R., Zonderman, A. B., Costa Jr, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory:

- Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, 70(3), 552.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30. <https://doi.org/1082-989X/96/S3.00>
- Menninghaus, W., Bohrn, I. C., Altmann, U., Lubrich, O., & Jacobs, A. M. (2014). Sounds funny? Humor effects of phonological and prosodic figures of speech. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1), 71. <https://doi.org/10.1037/a0035309>
- Menninghaus, W., Wagner, V., Wassiliwizky, E., Schindler, I., Hanich, J., Jacobsen, T., & Koelsch, S. (2018). What are aesthetic emotions? *Psychological Review*, 126(2), 171-195. <https://doi.org/10.1037/rev0000135>
- Mersch, D. (2019). Ästhetisches Denken: Kunst als Theoria [Aesthetic thinking: Art as theoria]. In D. Mersch, S. Sasse, & S. Zanetti (Eds.), *Ästhetische Theorie* [Aesthetic theory] (pp. 241-259). Diaphanes.
- Mills, A. (2003). The diagnostic drawing series. In C. Malchiodi (Ed.), *The handbook of art therapy* (p. 401-209).
- Mohammad, S. M., & Kiritchenko, S. (2018). *Wikiart emotions: An annotated dataset of emotions evoked by art*. Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).
- Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire design] (2nd ed.). Springer.
- Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-item HEXACO Persönlichkeitsinventars [Factor structure, psychometric characteristics and measurement invariance of the German version of the 60-item HEXACO personality inventory]. *Diagnostica*. <https://doi.org/10.1026/0012-1924/a000112>
- Mosquera, G. (1992). The Marco Polo syndrome: Some problems around art and Eurocentrism. *Third Text*, 6(21), 35-41. <https://doi.org/10.1080/09528829208576382>
- Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts*, 28(2), 173-191. <https://doi.org/10.2190/EM.28.2.d>
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396-402.

- O'Connor, B. P. (2018). SPSS, SAS, and MATLAB programs for determining the number of components and factors using parallel analysis and velicer's MAP test. <https://people.ok.ubc.ca/briocconn/nfactors/nfactors.html>
- Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, 9(11), 8.
- Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*, 12(2), 131-146.
- Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me... delight my eyes and brain: The Vienna integrated model of top-down and bottom-up processes in art perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. *Physics of Life Reviews*, 21, 80-125. <https://doi.org/10.1016/j.plrev.2017.02.003>
- Penn Center for Neuroaesthetics. (2019). Resources. <https://neuroaesthetics.med.upenn.edu/research.html> - resources
- Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4), 353-383.
- Raykov, T., & Little, T. D. (1999). A note on procrustean rotation in exploratory factor analysis: A computer intensive approach to goodness-of-fit evaluation. *Educational and Psychological Measurement*, 59(1), 47-57.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571.
- Schoch, K. (2014, February 25). *Ratinginstrument für zweidimensionale bildnerische Arbeiten: Methodische Arbeit zur Entwicklung eines psychologischen Tests zur Erfassung von bildnerischem Ausdruck [Rating instrument for two-dimensional pictorial works: Methodical development of a psychological test measuring pictorial expression]*. University of Mannheim. Zenodo database. <http://doi.org/10.5281/zenodo.1479744>
- Schoch, K., Gruber, H., & Ostermann, T. (2017). Measuring art: Methodical development of a quantitative rating instrument measuring pictorial expression (RizbA). *The Arts in Psychotherapy*, 55, 73-79. <https://doi.org/10.1016/j.aip.2017.04.014>
- Schoch, K., & Ostermann, T. (under review). Psychometrics of art: Validation of RizbA, a quantitative rating instrument for pictorial expression.

- Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1), 1-10.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.
- Silvia, P. J. (2007). An introduction to multilevel modeling for research on the psychology of art and creativity. *Empirical Studies of the Arts*, 25(1), 1-20.
- Silvia, P. J., & Nusbaum, E. C. (2011). On personality and piloerection: Individual differences in aesthetic chills and other unusual aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 5(3), 208. <https://doi.org/0.1037/a0021914>
- Simon, H., & Verstegen, U. (2004). prometheus: Das verteilte digitale Bildarchiv für Forschung und Lehre. Neuartige Werkzeuge zur Bereitstellung von verteiltem Content für Wissenschaft und Forschung [prometheus: Digital image database for research and education. New tools for providing shared content for science and research]. *Historical Social Research/Historische Sozialforschung*, 247-257.
- Specht, S. M. (2007). Successive contrast effects for judgments of abstraction in artwork following minimal pre-exposure. *Empirical Studies of the Arts*, 25(1), 63-70.
- Streb, J. H. (1984). Thoughts on phenomenology, education, and art. *Studies in Art Education*, 25(3), 159-166. <https://doi.org/10.1080/00393541.1984.11650370>
- Stuhler-Bauer, A., & Elbing, U. (2003). Die phänomenologische Bilderfassung: Ein kunsttherapeutisches Instrument. [The phenomenological picture survey: An art therapeutic instrument]. *Musik-, Tanz- und Kunsttherapie*, 14, 32-46. <https://doi.org/10.1026/0933-6885.14.1.32>
- Suhr, D. D. (2005). Principal Component Analysis vs. Exploratory Factor Analysis. *Proceedings of the thirtieth annual SAS® users group international conference*, 203, 30.
- Suhr, D. D. (2006). Exploratory or confirmatory factor analysis? *Statistics and Data Analysis*, 31.
- Sullivan, P., & McCarthy, J. (2009). An experiential account of the psychology of art. *Psychology of Aesthetics, Creativity, and the Arts*, 3(3), 181. <https://doi.org/10.1037/a0014292>
- Tinio, P. P. (2013). From artistic creation to aesthetic reception: The mirror model of art. *Psychology of Aesthetics, Creativity, and the Arts*, 7(3), 265-275. <https://doi.org/10.1037/a0030872>
- Tröndle, M., & Tschacher, W. (2012). The physiology of phenomenology: The effects of artworks. *Empirical Studies of the Arts*, 30(1), 75-113.

- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10.
- Walter, S., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1), 101-110.
- Wassiliwizky, E., Wagner, V., Jacobsen, T., & Menninghaus, W. (2015). Art-elicited chills indicate states of being moved. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4), 405. <https://doi.org/10.1037/aca0000023>
- Wünsch, K. L. (2016). Comparing two groups' factor structures: Pearson r and the coefficient of congruence. <http://core.ecu.edu/psyc/wuenschk/MV/FA/FactorStructure-TwoGroups.docx>
- Wundt, W. M. (1874). *Grundzüge der physiologischen Psychologie* [Characteristics of physiological psychology] (Vol. 1). Engelman.
- Zaidel, D. W., Nadal, M., Flexas, A., & Munar, E. (2013). An evolutionary approach to art and aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, 7(1), 100. <https://doi.org/10.1037/a0028797>
- Zumbo, B. D., Sireci, S. G., & Hambleton, R. K. (2003). *Re-visiting exploratory methods for construct comparability: Is there something to be gained from the ways of old*. Paper presented at the Annual Meeting of the National Council for Measurement in Education (NCME), Chicago, Illinois.

APPENDIX

Ethics approval

The authors received a positive vote of the Ethics Committee for Creative Arts Therapies at the Nürtingen-Geislingen University, Germany (Reference number 19C27J101W).

Funding

The study was funded by the Open Science Fellows Program by Wikimedia Germany, Stifterverband and VolkswagenStiftung (https://en.wikiversity.org/wiki/Wikimedia_Deutschland/Open_Science_Fellows_Program).

Abbreviations

CFA Confirmatory factor analysis

PCA Principal component analysis

PhP Hypertext Preprocessor

RizbA Ratinginstrument für zweidimensionale bildnerische Arbeiten (Rating instrument for two-dimensional pictorial works)

Additional Material

RizbA Original version

Table 7
***RizbA*: Original German version**

No.	Item
1	Das Bild enthält zeichnerische Elemente
2	Das Bild enthält malerische Elemente
3	Die Darstellungsweise ist gegenständlich
4	Die Darstellungsweise ist abstrakt
5	Der Farbauftrag ist pastos
6	Die vorherrschende Farbgebung ist leuchtend
7	Im Bild befinden sich vorwiegend reine Farben
8	Im Bild befinden sich vorwiegend Mischfarben (Sekundärfarben)
9	Im Bild sind Komplementärkontraste vorhanden
10	Im Bild enthaltene Formen sind vorwiegend organisch
11	Im Bild enthaltene Formen sind vorwiegend geometrisch
12	Die Linienführung verläuft vorwiegend gebogen
13	Die Linienführung verläuft vorwiegend eckig
14	Das Bild enthält unbearbeitete Flächen
15	Das Bild wirkt tief
16	Das Bild ist perspektivisch
17	Das Bild ist frei von Perspektive (aperspektivisch)
18	Das Bild ist unruhig
19	Das Bild ist wild
20	Die Gesamtkomposition ist senkrecht angelegt
21	Die Gesamtkomposition ist waagrecht angelegt
22	Die Gesamtkomposition ist diagonal angelegt
23	Die Gesamtkomposition ist flächendeckend ohne Hauptmotiv (All-Over-Structure)
24	Das Bild wirkt diffus
25	Das Bild wirkt präzise, exakt
26	Das Bild wirkt harmonisch

Open Data and Methodology

The list of artworks including metadata, the *SoSci Survey* structure (project structure, PHP code for randomization and drawing, values, variables), as well as the syntax for data structuring (Excel VBA macros) and statistical analysis (SPSS syntax) are available under a CC-BY 4.0 license at Open Science Framework via https://osf.io/p84xw/?view_only=8caa81c8ac954bb9ac7f56c4e5d2fe50.

A paper-pencil-version of the current original German version of *RizbA* questionnaire is available under a CC-BY 4.0 license at Zenodo:

Schoch, K. (2020, April 24). Ratinginstrument für zweidimensionale bildnerische Arbeiten (RizbA): Fragebogen mit Erläuterungen in deutscher Sprache [Rating instrument for two-dimensional pictorial works (RizbA): Questionnaire with explanatory notes in German language]. Zenodo. <http://doi.org/10.5281/zenodo.2530859>

Corresponding author at:

Kerstin Schoch, Institute for Art Therapy and Research, University of Applied Sciences and Arts Ottersberg, Am Wiestebruch 68, 28870 Ottersberg, Germany

E-mail: kerstin.schoch@hks-ottersberg.de

Thomas Ostermann, Department of Psychology and Psychotherapy, Faculty of Health, Witten/ Herdecke University, Alfred-Herrhausen-Straße 50, 58448 Witten, Germany

E-mail: thomas.ostermann@uni-wh.de