Empirics vs. art theory:
Exploring a factor structure of pictorial expression
based on contemporary artworks

Kerstin Schoch^{1,2}, Thomas Ostermann²

¹ HKS – University of Applied Sciences and Arts Ottersberg, Germany, ² Department of Psychology and Psychotherapy, Faculty of Health, Witten/ Herdecke University, Germany

Kerstin Schoch

HKS – University of Applied Sciences and Arts Ottersberg

Am Wiestebruch 68, 28870 Ottersberg, Germany

kerstin.schoch@hks-ottersberg.de

(corresponding author)

Thomas Ostermann

Department of Psychology and Psychotherapy

Faculty of Health, Witten/ Herdecke University

Alfred-Herrhausen-Straße 50, 58448 Witten, Germany

thomas.ostermann@uni-wh.de

Abstract

As a transdisciplinary questionnaire combining psychometrics and art theory, RizbA enables a quantitative measurement of pictorial expression in terms of a formal image analysis. This study explores its factor structure and a potential gap between theory and empirics.

The sample consists of 275 contemporary pictorial works by artists and nonprofessionals, rated by 179 experts in a randomized online survey. Three path models were specified and computed as CFA: models A and B based on the results of the two previous studies, model C on the theoretical framework of the initial study. Model C was additionally tested on a combined dataset of all three studies.

While models A and B did not converge, model C was associated with fit indices as follows: $\chi^2 = 1299.752$, $df = 278$, $p = .000$, RSMEA = .122 (90% CI = .116, .129), CFI = .712, TLI = .679, SRMR = .135 and for the combined dataset: $\chi^2 = 6860.824$, $df = 278$, $p = .000$, RSMEA = .086 (90% CI = .084, .088), CFI = .740, TLI = .696, SRMR = .084.

Studies implicate reliability, but there might not be a globally stable factor structure across artworks. Only model C partly suggests an acceptable fit for the combined data. The results speak to a methodological gap between empirics and theory due to art being highly ambiguous with various analysis options. Further postdisciplinary approaches are needed to develop a theory-based measurement model for pictorial expression, which might do justice to the arts.

Keywords: art theory, confirmatory factor analysis, formal picture analysis, pictorial expression, visual art

Introduction

Psychological research on visual art has a long tradition. Surprisingly, however, until now formal elements of artworks have received little attention in quantitative research. Another peculiarity is that most studies deal with historical art, but seldom with contemporary works. This is even more striking since arts and humanities provide a range of methods to analyze artworks including contemporary art (e.g., Bockemühl, 1989; Huber, 2005; Kemp, 1988; Panofsky, 2006). In order to do justice to the arts, approaches from empirical sciences and humanities will need to move towards becoming more transdisciplinarily connected and share their discourses, and this study contributes to that process.

As already discussed in detail (Schoch et al., 2017; Schoch & Ostermann, 2020, 2021, May 4), RizbA (Rating instrument for two-dimensional pictorial works) fills the severe lack of a quantitative, reliable and validated instrument for conducting a detailed formal analysis of artworks, that allows inferential statistics and is applicable to all sorts of two-dimensional pictorial works.

The rating instrument for two-dimensional pictorial works (RizbA)

RizbA is an Open Methodology tool. The 26-item questionnaire (see Table 1) refers to a formal picture analysis and focuses on a maximization of objectivity. The instrument uses a six-point Likert-scale, which is discretely scaled and verbally anchored in shades of agreement (0 = *strongly disagree* to 5 = *strongly agree*). Raters receive a brief instruction to process the presented image: They are asked to focus on the pre-dominant overall expression and no single details, while being reminded that there is no right or wrong, but only their evaluation. Explanatory notes explaining art terminology ensure that the questionnaire remains applicable and reliable (Jerusalem, 2020, February 20) not only to art experts, but also to experts from other domains.

Table 1*Rating instrument for two-dimensional pictorial works (RizbA)*

No.	Item*
1	The picture includes graphic elements
2	The picture includes pictorial elements
3	The manner of representation is concrete
4	The manner of representation is abstract
5	The color application is pastose
6	The predominant coloring is vibrant
7	In the picture primary colors are prevalent
8	In the picture mixed colors (secondary colors) are prevalent
9	In the picture there are complementary contrasts
10	In the picture organic shapes are prevalent
11	In the picture geometric shapes are prevalent
12	The layout of the line is predominantly curved
13	The layout of the line is predominantly angled
14	The picture includes unworked areas
15	The picture appears to be deep
16	The picture is perspectival
17	The picture is without perspective (aperspectival)
18	The picture is restless
19	The picture is wild
20	The global composition is laid out vertically
21	The global composition is laid out horizontally
22	The global composition is laid out diagonally
23	The global composition is laid out area-wide without a main subject (All-Over-Structure)
24	The picture appears to be diffuse
25	The picture appears to be precise, accurate
26	The picture appears to be harmonic

* = original German version see additional material

Pictorial expression

Defining the nomological network, RizbA assesses the construct of pictorial expression, which is defined as artistic creations in the form of a picture. It focuses on the concept of a formal picture analysis (Streb, 1984; Stuhler-Bauer & Elbing, 2003) originating from the humanities and the arts. It analyzes formal aspects, such as representation, color, spatiality, shaping, pictorial elements and composition, which can be found across art literature (e.g. Arnheim, 2013; Bauer, 1996; Kandinsky, 1955; Meyer, 2011; Vollmar, 2008). This approach is rooted in the tradition of phenomenological picture analysis (Stuhler-Bauer & Elbing, 2003, 2004). In art therapy, the phenomenological method consists of attempting attentive viewing and describing as precisely as possible, what is seen (Betensky, 1991). The viewer seeks to overcome accidental judgement, preconception and association (Streb, 1984). The relevant systemic level is the pictorial representation consciously omitting the picture reference, e.g. knowledge,

associations, emotions (Huber, 2005). It focuses on visual presentation (Wiesing, 2005) leaving out the colonializing view of unintentional identification of objects (Marotzki & Stoetzer, 2006). It is limited to a detailed but classical conception of images, not taking into account the creation process (Uzelac, 1998). The test does not judge the creator's achievement or mastery and is distinct from aesthetic appreciation. It is neither evaluative nor interpretative nor projective but aims for a value-free description of formal elements. A rater training or sample images for certain characteristics are deliberately dispensed with to avoid a manipulation of judgement.

Previous research

So far three main validation studies have been conducted using randomized online surveys, in which image samples were rated by experts from visual art disciplines. Focusing on statistical quality criteria, these studies successfully validated the questionnaire in terms of item difficulty, capacity for differentiation, test-retest and inter-rater reliability. Results indicate good to high reliabilities and allow a generalizability to professional and nonprofessional contemporary art. Still, the question of the factor structure remained partly unsolved.

In an initial study (Schoch et al., 2017), RizbA was constructed based on a theoretical framework of content areas and empirically tested on a small heterogeneous sample of pictorial works by nonprofessionals. A pool of 113 items (Schoch, 2014, February 25) was gathered based on art literature, compendia (e.g., Arnheim, 2013; Kandinsky, 1955; Meyer, 2011; Vollmar, 2008) and art therapeutic questionnaires (e.g., Eber et al., 1998; Elbing & Hacking, 2001; Hacking et al., 1996; Stuhler-Bauer & Elbing, 2004). Based on statistical quality criteria it was narrowed down to 26 items. The underlying measurement model was thought of in terms of a formative model. Pictorial expression was – and still is – not seen as a one-dimensional latent variable, but rather a profile of characteristics which complement to the construct. Since there was no quantitative data on formal image analysis by then, *a priori* creating a construct

map was hardly possible. A second study (Schoch & Ostermann, 2021, May 4) validated the 26-item version on a sample of 294 pictorial works by nonprofessionals from the 21st century. . A principal component analysis (PCA) resulted in eight factors (see Table 2). In a third study, the instrument was validated on a sample of 318 works dated to the 21st century by professional artists (Schoch & Ostermann, 2020). A PCA also resulted in eight factors which differed slightly from the study (see Table 2).

Seen across the studies, the theoretical framework of content areas from the first study and the PCA results from the second and the third studies differed substantially. The validation studies included iterative discussion processes on art theory with expert groups. Those came to the intriguing conclusion that the empirical PCA solutions do not appropriately reflect art-theoretical categories.

Goal of this study

Previous studies computed a PCA as an exploratory procedure for data reduction to derive a number of components that account for the variability in measures (DeCoster, 1998). This study makes a further inquiry into an underlying factor structure. With reference to the previous three studies, three different models are hypothesized, specified and empirically tested on another dataset – either based on empirical results or art theory. As a hypothesis-testing procedure, a CFA was conducted to test if there is a sufficient model fit between each hypothesized model and the data (Moosbrugger & Kelava, 2007; Suhr, 2005, 2006). To examine the factors' quality, statistical criteria such as descriptive statistics, normal distribution, internal consistency and intercorrelations were calculated.

Material and methods

Pictorial material

Referring to previous studies and a *ceteris paribus* assumption, the pictorial stimulus material consisted of contemporary pictorial works by professional artists and nonprofessionals. The

required sample size of images was *a priori* calculated based on statistical guidelines. The sample size regarding images is 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.

Artworks by professional artists

The image sample of visual artworks by professional artists was generated using WikiArt (2021) images under a Creative Commons License. For the systematic search and retrieval of images and metadata, the Open Source API for WikiArt (2021) and WikiArt Retriever (Davis, 2018) were used. The data was retrieved in June 2020.

The inclusion criteria for images were determined analogously to the third study as follows: Firstly, as a criterion for being contemporary, a creation date of between 2018 and 2020 had to be specified in the metadata. The WikiArt Retriever (Davis, 2018) code was accordingly modified to retrieve only these works. Secondly, only handmade techniques such as drawings, paintings, collages and mixed techniques were used. Thirdly, the image had to comprise a complete view. Images showing only a detail of an artwork or an entire exhibition space displaying the artwork were excluded. The resulting sample was reduced to one image per artist, choosing the most recent work of that artist. If several works were dated to the same year, the first one, following alphabetical order was selected.

The final sample of professional contemporary artworks consisted of 143 images. Images, in which a background around the actual work was to be seen, were edited using Photos for MacOS (Apple, 2020) and cropped down to show solely the artwork.

Artworks by nonprofessionals

Nonprofessional artworks were collected via a public call, in which nonprofessionals were invited to digitally submit their works. The data was collected during 2020.

Inclusion criteria were determined as follows: Firstly, participants were asked to participate only if they had no professional education in visual arts or previous experience of studying a subject related to visual arts. Secondly, referring to previous studies and *a ceteris paribus* assumption only handmade techniques such as drawings, paintings, collages and mixed techniques were included.

The call was disseminated via a website, email, social media and particularly in specific Facebook groups with potential suitable participants. For finding such groups, the keywords *visual art* were searched in various languages using DeepL (2020) translator. Additionally, a large set of other keywords were used in English and German language, such as *amateur art*, *hobby art*, *painting*, *drawing*, *international art*, *Black art/ists*, *African American art/ists*, *native American art/ists*, *Korean art/ists*, *Armenian art/ists*, *Vietnamese art/ists*, *indigenous art/ists*, *Asian art/ists*, *Arabic art/ists*, *Persian art/ists*, *Islamic art/ists*, *Indian art/ists*, *Zambian art/ists*, *South African art/ists* etc.

Potential participants could anonymously submit their image via an upload form using Wordpress (2020) and the plugin Contact Form 7 (Miyoshi, 2020). The instructions were as follows: First, participants were invited to create a two-dimensional picture. They could also use a picture they had already created. Participants could decide on material and size. It was noted that making a submission was not a painting competition and not about creating a particularly "beautiful" picture, but simply about creating. Second, they were asked to take a photo of the picture using a smartphone or digital camera and were provided with tips and photo examples for taking a high-quality picture. Third, they were invited to upload their photo accompanied with information about the picture's size and artistic material, demographic data and informed consent.

A total of 139 images were gathered this way. Seven of these were excluded due to not fitting the inclusion criteria or lacking image quality (e.g., resolution, sharpness, picture section). The

final sample consisted of 132 images by 132 nonprofessionals (105 female, 25 male, 2 diverse) from 21 to 79 years old ($M = 45.80$, $SD = 15.12$) with a highest educational qualification ranging from secondary school to PhD. Where necessary, images were edited using MacOS (Apple, 2020) and the Open Source image editor Gimp (Neumann, 2020). Those, in which a background around the actual work was visible, were cropped down to solely the artwork. Distorted perspective due to a skewed photography was adjusted. A few were brightened up.

Final image sample

The final sample ($N = 275$) consists of 143 contemporary artworks by professional artists and 132 pictorial works by nonprofessionals. Both subsamples include international image material from different geographic areas.

Study

Study design

An online study design was chosen using SoSci Survey (Leiner, 2018) with the additional use of PHP elements. In this survey, experts with a professional training related to visual art were asked to rate images using RizbA items. Each rater was presented with six images, randomly chosen from the image sample. Raters were recruited via email, websites and social media. Among others, the call was posted in 132 social media groups and sent via email to 123 universities and 47 associations in Germany, Austria and Switzerland. The latter were contacted and their professional staff, students and associates were invited to participate and disseminate the call. No rater training was conducted. As an incentive, vouchers were raffled among participants.

Raters

As a criterion for inclusion and in order to ensure raters were aware of the appropriate terminology, only professionals who had an academic degree related to visual art, or were students who had been studying an art-related subject for at least one year were requested to

participate. Relevant disciplines included art history, art pedagogy, visual art, art therapy, design, graphic design, art sciences, image sciences, restoration and others. Regarding the sample size of raters Walter et al. (1998) recommend two or three observations per subject, given $\alpha = .05$ and a required reliability of at least .40. Based on these parameters the current study aimed at obtaining three raters per picture. Taking into account that each rater had to process six images, this led to an *a priori* calculated sample size of at least 138 raters.

Procedure

All images were reduced to a file size of 21-243 KB using Apple Photos software to adjust them to SoSci Survey specifications. In the online survey, raters were introduced to the scale and asked to rate each picture consecutively presented to them using all 26 items. For each rater, six images were randomly, successively drawn from a pot, without putting them back. When all the images from the entire picture sample had been rated once, the process started again. The order in which the items were presented was randomized between subjects. Raters could not skip items but were obliged to answer all questions in order to proceed. At the end raters were asked if they had completed the questionnaire carefully.

Statistical Analysis

The data was analyzed using R version 4.0.3 and R Studio version 1.3.1093 including the packages psych (Revelle, 2016), readxl (Wickham & Jennifer Bryan, 2013), car (Fox et al., 2019), lavaan (Rosseel, 2020), pbivnorm (Genz & Kenkel, 2015), and semPlot (Epskamp et al., 2019). Only complete datasets were used, in which all six images presented were fully processed. Data from prematurely terminated surveys or from raters who stated that they had not completed the questionnaire carefully were excluded. Because each rater assessed only six images out of the entire collection, the dataset inherently included a large number of empty cells. For analyses regarding the items and factors, subjects were therefore merged to new,

generic combined raters. Descriptive statistics were calculated for the sample of raters in terms of demographic variables, such as self-identified gender, age and art-related qualification.

Three different path models were specified as hypothetical models. Items were accordingly assigned to the factors (see Table 2):

Model A is empirically based on the second study (Schoch & Ostermann, 2021, May 4) (Schoch & Ostermann, 2021, May 4) resembling the factor structure from the PCA at T1. Items with negative factor loadings were reversed in the model.

Model B is empirically based on the third study (Schoch & Ostermann, 2020) resembling the factor structure from the PCA at T1. Items with negative factor loadings were reversed in the model.

Model C is theoretically-based on the initial study (Schoch et al., 2017) and resembles the framework of content areas that was *a priori* used for compiling the initial item pool. In comparison to the content areas (*representation, color, shape, space, motion, composition, expression*) the factor labels were adjusted slightly (see Table 2) to reflect the insights gained by expert group evaluations on the tool. No items were reversed.

For each factor of each model the following analyses were conducted: Descriptive statistics and Shapiro-Wilk normality test and Cronbach's alpha were computed. For the latter interpretation was based on Gliem and Gliem (2003) describing an alpha of .80 as a reasonable goal and George and Mallery (2003) suggesting the following rules of thumb: $\geq .90$ excellent, $\geq .80$ good, $\geq .70$ acceptable, $\geq .60$ questionable, $\geq .50$ poor, and $< .50$ unacceptable. For each model, Pearson product moment correlations between factors were calculated and interpreted based on the usual conventions: .10 small, .30 medium, and .50 large (Cohen, 1992). For inferential statistics the α level was *a priori* defined to be .05.

For each model a CFA was conducted using the model syntax by Rosseel (2021). With the factors being latent variables in the regression formula, they were defined by listing their manifest indicators, the items they are hypothesized to be measured by. For each factor in models A and B the items were ordered from highest to lowest factor loading, based on previous PCA results. As recommended by Li (2016) when the normality assumption is slightly or moderately violated, the robust maximum likelihood estimator (MLR) with robust standard errors and a Satorra-Bentler scaled test statistic (Rosseel, 2021) was used.

For model C an additional CFA was computed on a combined dataset of both the current and the two previous studies, resulting in a total of 894 images rated with RizbA. Models A and B were not tested on the previous data, since this would result in a deceptively optimistic overfitting (Fokkema & Greiff, 2017) and be tautological.

The adequacy of models fit was evaluated using indices as recommended in the corresponding literature (Hu & Bentler, 1999; Lorenzo-Seva & Ten Berge, 2006; Moosbrugger & Kelava, 2007; Parry, 2017; Sen et al., 2014; Suhr, 2006; Tucker & Lewis, 1973): χ^2 statistics in relation to the degrees of freedom (*df*), the comparative fit index (CFI), the Tucker-Lewis Index (TLI), the root-mean-square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR). Due to the normality assumption being slightly violated for some factors, the robust version of these indices was used.

We oriented towards the following recommendations on cutoff values to be interpreted as a fit between the hypothesized model and the observed data: Moosbrugger and Kelava (2007) recommend, that χ^2/df values between .000 - 2.00 suggest a good and those between 2.01 and 3.00 an acceptable fit. CFI values between .970 and 1.00 imply a good and those between .950 and .969 an acceptable fit. RMSEA values between .000 and .050 can be interpreted as a good, values between .051 and .080 as an acceptable fit. Hu and Bentler (1999) recommend for the ML method cutoff values $\geq .95$ for CFI and TLI, $\leq .06$ for RMSEA – according to Shi et al.

(2019) a RSMEA $\geq .10$ is beyond consideration – and close to .08 for SRMR. Concerning TLI Lorenzo-Seva and Ten Berge (2006) suggest, that values in a range of .85 - .94 correspond to a fair similarity, while values $\geq .95$ imply that the two components compared can be considered equal.

Table 2*Models A, B, and C: Assigned items*

Model A		Model B		Model C	
empirically based		empirically based		theory-based	
(Schoch & Ostermann, 2021, May 4)		Schoch and Ostermann (2020)		Schoch et al. (2017)	
Factor label	Items	Factor label	Items	Factor label	Items
<i>Picture effect</i>	18, 19, 24, 25 _r , 26 _r	<i>Picture effect</i>	18, 19, 24, 25 _r , 26 _r	<i>Representation</i>	1, 2, 3, 4, 5
<i>Spatiality</i>	15, 16, 17 _r	<i>Spatiality</i>	15, 16, 17 _r	<i>Color</i>	6, 7, 8, 9
<i>Shaping</i>	10, 11 _r , 12, 13 _r	<i>Shaping</i>	10, 11 _r , 12, 13 _r	<i>Shaping</i>	10, 11, 12, 13, 14
<i>Pictorial elements</i> (drawing vs. painting)	1, 2 _r , 5 _r , 14	<i>Pictorial elements (painting)</i>	2, 5, 7 _r , 8	<i>Spatiality</i>	15, 16, 17
		<i>Pictorial elements (drawing)</i>	1, 14, 22	<i>Motion</i>	18, 19
<i>Representation</i>	3 _r , 4, 22, 23	<i>Representation</i>	3 _r , 4, 23	<i>Composition</i>	20, 21, 22, 23
<i>Color intensity</i>	6, 9	<i>Color intensity</i>	6, 9	<i>Picture effect</i>	24, 25, 26
<i>Color mixture</i>	7 _r , 8				
<i>Composition</i>	20, 21 _r	<i>Composition</i>	20, 21 _r		

_r = reversed item with negative factor loading in the previous PCA

Results

Raters

A total of 184 raters finished the survey. Data from one rater was excluded from the analysis after they stated that they had not filled out the survey carefully carefully. Data from four more raters were excluded, because they did not meet the inclusion criteria of expertise. The final sample of valid raters consists of 179 experts (159 female, 16 male, 4 diverse) between 22 and 72 years ($M = 37.59$, $SD = 12.51$). All of them have an academic degree in an art-related subject (75.42 %) or are at least in their second year of study (24.58 %). The disciplines include art therapy (31.84 %), art education (15.64 %), art history (13.41 %), fine arts (8.94 %), art science (5.59 %), graphic design (5.59 %), restoration (5.03 %), design (3.35 %), and others (10.61 %). Each picture is rated by 0 to 7 raters ($M = 3.91$, $SD = 1.39$). One nonprofessional's artwork was unrated and therefore excluded from further analysis.

Factors

The descriptive statistics are reported in Table 3, Table 4, and Table 5. Shapiro-Wilk test yields significant p-values in six of eight factors in model A and B, and in four of seven in model C. Cronbach's alpha ranges between .39 and .91 (see Table 3, Table 4, Table 5). Intercorrelations between the factors are reported in Table 6, Table 7, and Table 8.

Table 3*Model A: Descriptive statistics, Shapiro-Wilk test and Cronbach's alpha*

Factor	Factor label	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	Shapiro-Wilk		Cronbach's α		
						<i>W</i>	<i>p</i>	<i>r</i>	95% <i>CI</i>	
									<i>lower bound</i>	<i>upper bound</i>
1	<i>Picture effect</i>	2.15	0.81	0.40	4.60	0.99	.028	.88	.86	.90
2	<i>Spatiality</i>	2.56	1.06	0.00	4.89	0.99	.030	.91	.89	.91
3	<i>Shaping</i>	3.02	0.94	0.00	4.75	0.94	.000	.88	.86	.91
4	<i>Pictorial elements</i> (drawing vs. painting)	2.30	0.85	0.25	4.85	0.97	.000	.70	.64	.75
5	<i>Representation</i>	1.94	0.91	0.25	4.50	0.96	.000	.73	.69	.78
6	<i>Color intensity</i>	2.47	1.05	0.00	4.50	0.98	.001	.68	.61	.68
7	<i>Color mixture</i>	2.95	0.79	0.50	4.75	0.98	.000	.39	.25	.53
8	<i>Composition</i>	2.74	1.07	0.50	5.00	0.97	.000	.76	.71	.82

M = mean, *SD* = standard deviation, *Min* = minimum, *Max* = maximum, *W* = Shapiro-Wilk test statistic, *p* = p-value Shapiro-Wilk test, *r* = reliability coefficient Cronbach's alpha, *CI* = confidence interval

Table 4*Model B: Descriptive statistics, Shapiro-Wilk test and Cronbach's alpha*

Factor	Factor label	M	SD	Min	Max	Shapiro-Wilk		Cronbach's α		
						W	p	r	95% CI	
									lower bound	upper bound
1	Picture effect	2.15	0.81	0.40	4.60	0.99	.028	.88	.86	.90
2	Spatiality	2.56	1.06	0.00	4.89	0.99	.030	.91	.80	.93
3	Shaping	3.02	0.94	0.00	4.75	0.94	.000	.88	.86	.91
4	Pictorial elements (painting)	2.78	0.70	0.81	4.31	0.98	.000	.61	.54	.68
5	Pictorial elements (drawing)	3.04	0.79	0.17	4.00	1.00	.000	.39	.27	.51
6	Representation	2.01	1.19	0.00	4.67	0.93	.000	.88	.85	.90
7	Color intensity	2.47	1.05	0.00	4.50	0.98	.001	.68	.61	.76
8	Composition	2.74	1.07	0.50	5.00	0.97	.000	.76	.71	.82

M = mean, *SD* = standard deviation, *Min* = minimum, *Max* = maximum, *W* = Shapiro-Wilk test statistic, *p* = p-value Shapiro-Wilk test, *r* = reliability coefficient Cronbach's alpha, *CI* = confidence interval

Table 5*Model C: Descriptive statistics, Shapiro-Wilk test and Cronbach's alpha*

Factor	Factor label	<i>M</i>	<i>SD</i>	Min	Max	Shapiro-Wilk		<i>r</i>	Cronbach's α	
						<i>W</i>	<i>p</i>		95% CI	
									lower bound	upper bound
1	<i>Representation</i>	2.66	0.42	0.42	1.00	0.99	.010	.63	.56	.63
2	<i>Color</i>	2.40	0.76	0.08	4.00	0.95	.000	.62	.55	.62
3	<i>Shaping</i>	2.14	0.40	0.93	3.24	1.00	.754	.77	.72	.77
4	<i>Spatiality</i>	2.33	0.40	0.38	1.33	0.99	.472	.91	.89	.91
5	<i>Motion</i>	2.18	1.02	0.00	5.00	0.99	.009	.92	.90	.92
6	<i>Composition</i>	1.98	0.41	0.25	3.19	0.98	.002	.49	.40	.49
7	<i>Picture effect</i>	2.53	0.38	1.00	3.67	0.99	.010	.75	.70	.75

M = mean, *SD* = standard deviation, Min = minimum, Max = maximum, *W* = Shapiro-Wilk test statistic, *p* = p-value Shapiro-Wilk test, *r* = reliability coefficient Cronbach's alpha, CI = confidence interval

Table 6*Model A: Pearson product moment correlations between factors*

	1	2	3	4	5	6	7	8
1 Picture effect	—							
2 Spatiality	-.37**	—						
3 Shaping	-.05	.13	—					
4 Pictorial elements (drawing vs. painting)	-.05	-.18*	-.04	—				
5 Representation	.42**	-.46**	-.27**	-.16	—			
6 Color intensity	.09	.05	.03	-.43**	.12	—		
7 Color mixture	.02	.08	-.07	-.33**	.03	.00	—	
8 Composition	-.06	-.10	.06	.02	-.15	-.09	.02	—

* $p < .05$. ** $p < .01$ **Table 7***Model B: Pearson product moment correlations between factors*

	1	2	3	4	5	6	7	8
1 Picture effect	—							
2 Spatiality	-.37**	—						
3 Shaping	-.05	.13*	—					
4 Pictorial elements (painting)	.06	.18*	.10	—				
5 Pictorial elements (drawing)	-.05	-.02*	.10	-.36**	—			
6 Representation	.45**	-.53**	-.27**	.02	-.22**	—		
7 Color intensity	.09	.05	.03	.25**	-.29**	.11	—	
8 Composition	-.06	-.10	.06	.02	-.04	-.11	-.09	—

* $p < .05$. ** $p < .01$ **Table 8***Model C: Pearson product moment correlations between factors*

	1	2	3	4	5	6	7
1 Representation	—						
2 Color	0.29	—					
3 Shaping	0.00	-0.17	—				
4 Spatiality	0.12	0.12	-0.17	—			
5 Motion	0.10	0.13	-0.02	-0.14	—		
6 Composition	0.01	0.18	-0.05	0.20	0.16	—	
7 Picture effect	-0.03	0.11	0.07	0.33	-0.30*	0.11	—

* $p < .05$. ** $p < .01$ **CFA**

Models A and B did not converge. Model C (see **Error! Reference source not found.**) revealed the following fit indices: $\chi^2 = 1299.752$, $df = 278$, $p = .000$, RMSEA = .122 (90% CI = .116, .129), CFI = .712, TLI = .679, SRMR = .135 and for the combined dataset of three studies as follows: $\chi^2 = 6860.824$, $df = 278$, $p = .000$, RMSEA = .086 (90% CI = .084, .088), CFI = .740, TLI = .696, SRMR = .084.

Figure 1 CFA Path model C with intercorrelations between factors

Discussion

Previous studies have aimed at exploring the underlying factor structure of RizbA. This paper aims to compare the existing two empirical-based PCA factor models with a theoretical model by means of CFA path models. We were able to show substantial differences between these

approaches. Below, some methodological aspects will be discussed, followed by a broader discussion from a content and art related perspective.

Rater sample size

Preceding validation studies on RizbA used a large number of 10 to 20 raters per image and resulted in sufficiently high inter-rater agreement and test-retest reliability. This study uses a smaller rater sample. This is both more economic and also sufficient. First, because the instrument has been proven to be reliable in terms of inter-rater agreement and test-retest reliability. Second, the current investigation concentrates on artistic stimulus material rather than on intersubjectivity.

Normal distribution

The Shapiro-Wilk test suggests a violation of normal distribution for several factors for each model. However, it is a conservative procedure and there is no theoretical reasoning as to why pictorial expression should not be normally distributed within a larger sample size. For example, model A's factors *shaping*, *pictorial elements (drawing vs. painting)*, *color intensity*, *color mixture*, and *composition* are not distributed normally in this sample. One reason might be that all images were created on purpose, with the knowledge that they would be shown to others. This applies to the professional artworks uploaded to WikiArt, but also for those willingly contributed by nonprofessionals. This might result in a tendency towards more colored or elaborately shaped pictures, while the representation is still balanced between abstract and representational and therefore normally distributed. For testing this rationale, random pictures could be included, e.g., images that are not even considered to be artworks such as scribbles on slips of paper.

Internal consistency

Models A and B provide better internal consistency within factors than model C. For model A Cronbach's alpha can be interpreted as excellent for the factor *spatiality*, good for *picture*

effect and *shaping*, and acceptable for *pictorial elements* (*drawing* vs. *painting*), *representation* and *composition*. However, the reliability is to be seen questionable for *color intensity* and unacceptable for *color mixture*. Analogously, results from model B suggest excellent reliability for the factor *spatiality*. Also, it can be interpreted as good not only for *picture effect* and *shaping*, but also for *representation*. *Composition* yields acceptable, while *pictorial elements* (*painting*) and *color intensity* only show poor, and *pictorial elements* (*drawing*) even unacceptable reliability. Regarding model C, the factor *spatiality* again suggests excellent reliability along with *motion*. *Shaping* and *picture effect* are acceptable, *representation* and *color* questionable, while *composition* is unacceptable.

Intercorrelations

The following factors show intercorrelations: Within model A *spatiality* negatively correlates with *picture effect* and *pictorial elements* (*drawing* vs. *painting*). Representation is related to *picture effect* and negatively to *spatiality* and *shaping*. *Color intensity* and *color mixture* both negatively correlate with *pictorial elements* (*drawing* vs. *painting*). Within model B *spatiality* correlates with *shaping*, *pictorial elements* (*painting*), and *pictorial elements* (*drawing*) as well as negatively with *picture effect*. Representation is related with a variety of other factors: with *picture effect* and negatively with *spatiality*, *shaping* and *pictorial elements* (*drawing*). Also, *color intensity* is associated with *pictorial elements* (*painting*) and negatively with *pictorial elements* (*drawings*), which correlates to a painting being more likely to be color intensive than a drawing. Within model C factors are less related. Only *picture effect* and *motion* show medium correlations.

Factor structure

Models A and B do not provide model fit indices. For model C all indices yield values beneath the cut-off points associated with an acceptable fit. Only when tested on the combined data of all three studies, does SRMR suggest an acceptable to good model fit. However, none of the

hypothesized models provides a factor model with a thoroughly good or acceptable fit – neither empirically nor theory-based. As an exploratory approach we conducted an *a posteriori* card sorting test with art experts to generate further theory-based models. When tested with a CFA they converged – in comparison to the empirically based models A and B – but still did not provide a good model fit. It can be assumed, that the theory-based models are more likely to fit than the empirically based models, based on PCAs. There are a number of possible reasons, which we will discuss below.

First, a sample of 275 observations (images) – 894 in case of the combined dataset – and 26 variables (items) might still not be large enough to map a factor structure, since CFA is a large sample procedure. According to Shi et al. (2019) a sample of 200 observations only provides a reasonable estimate for CFI and TLI when the number of observed variables is less than 30.

Second, the CFA do not reflect results of preceding PCAs although the sample consisted of a similar sample. This might be due to differences in the calculation of both procedures. Also, the PCA solutions suppressed factor loadings under .40. This is a common practice and if these solutions were stable, a similar structure might be found in the CFA. The fact that these models do not converge, could indicate that the solution is not stable across art. This suggests that measurement of pictorial expression is even more extensive and complicated than anticipated and there might not be a stable global solution.

Third, there is a methodical gap between empirics and theory when it comes to formal image analysis. Pictorial expression in terms of a formal image analysis might be an even more complex construct than assumed. It is possible that there is no underlying factor structure analogous to psychological constructs, such as personality or intelligence. Or possibly the underlying factor is not stable across a larger range of artworks. To gain insight, we have to dive deeper in the various perspectives on image analysis, of which the perspectives addressed below represent only a glimpse of the field.

One established method for image analysis is the *iconographic-iconological method*, which Panofsky (1978, 2006) abstracted to a theoretical model (Eberlein, 2008). It suggests three stages: First, the *pre-iconographic description* determines the primary or natural subject and identifies the objects presented in an image. It requires a rudimentary knowledge of style history as a corrective principle for understanding the image objects as an artistic motive. As an example Eberlein (2008) uses Friedrich Dürer's *Melencolia I*, which today might be interpreted as a grumpy angel in the midst of stuff; however familiarity with the basic art principles from 1514 would anticipate a personification and a symbolic meaning of the objects. Second, the *iconographic analysis*, which uses historical sources and type history to unveil underlying themes and ideas. Third, the *iconological interpretation* examines the actual meaning and content, instrumentalizing the time difference between artwork and interpreter. It refers to Ernst Cassirer's philosophy of symbolic forms, representing the idea that humans no longer live in a merely natural universe, but experience the world via language, myth or religion, and art. If we consider *pre-iconographic* description not only in motives but also in shapes, there is a parallel to RizbA. However, the *iconographic-iconological method* decrypts images based on historical contextualization, implying that an artwork can never be detached from its historical context. RizbA leaves aside all historical aspects of the epochs artworks stem from, which might be one missing piece of the big picture when measuring art.

Another missing piece might be argued from the perspective of a *reception aesthetics* approach (Kemp, 1988), which suggests that every artwork provides a receptive function. The way art encounters the viewer depends on the way the viewer approaches the art: responding and acknowledging the viewer's effort. Reception aesthetics is oriented towards the artwork and is in search of an implicit viewer, a viewer function in an artwork. Each image is addressed to someone, framing its recipient, and discloses two sorts of information: It speaks about its place and impact in society. And it speaks about itself. Consequently, reception aesthetics has (at least) three tasks: Firstly, it has to recognize the means by which the artwork gets in touch with

us and read them with regard to secondly, their social-historical, and thirdly, their actual aesthetic statement. In summary, recipient and artwork are not clinically pure and isolated entities. Although RizbA does not explicitly address internal psychological processes, the items summarized as *picture effect* refer to those. However, the described lack of pure constructs might be another note on a potential non-existence of a stable factor solution.

RizbA is initially based on the framework of *phenomenological image analysis* seeking to overcome judgement, preconception and association (Streb, 1984), while being aware that there is no such thing as absolute objectivity. It is rather a process of reflecting on one's own perceptions. Modern phenomenology goes back to Edmund Husserl and is interested in phenomena (things, objects) and how they present themselves as perceived experiences (Betensky, 1991). Husserl postulates that an artwork possesses an invisible space that belongs to the picture and is opposed to the space, that can be grasped by experience (Uzelac, 1998). The part of an artwork that cannot be grasped might be another missing piece that may not be objectively measurable.

Finally, Bockemühl (1989) argues that analogies of the artwork as sender and the eye as receiver are too simplistic. Art makes the flow of information rather complicated and denies a final statement. It creates a reality of its own, in which only the combination of the creator's hand and the creative power of the eye lead to the artwork, e.g., Mark Rothko's colorfield paintings that slowly arise by looking at them. An image is an open system, that never offers an exhaustive interpretation. We have no choice but to accept certain limits and that analysis only captures a part – even though it is indispensable – of artistic reality. When dealing with art, one seems to be pushed into subjectivism. There might only be one escape: reducing the statement of art to a pure informational level and excluding further questions that might be – how as (Bockemühl, 1989) drastically states it – non-scienceable.

Limitations

The professional art images were retrieved from WikiArt (2021), a shared knowledge database. Unfortunately the accompanying metadata lacks the artists' self-identified gender or other diversity indicators (Pacher-Nishikawa et al., 2021). Thus, it was not possible to compute demographics to ensure heterogeneity and diversity of the image sample. The works by nonprofessionals were gathered via a call using social media. Open Science advocates justifiably criticize commercial platforms' non-transparent algorithms and problematic data protection. Nevertheless, in this study Facebook groups were used since they offer an efficient way to reach specific target groups. The search mechanisms do not provide a systematic search function. Its results are biased toward the user's language and region, hence towards English and German speakers.

In the online survey image sizes were rather small, in order to meet the SoSci Survey's (Leiner, 2018) maximum upload capacity and to make sure images load fast enough. Also, the raters viewed the images on their own devices, which might result in a different display of color and resolution. Using a higher resolution, a different rating experience may be expected, especially when original works instead of digital copies are used. This is particularly true if we are to follow art theories (e.g., Panofsky, 1978; Panofsky, 2006), who claim that an artwork cannot be completely comprehensible if it is detached from its context. Due to the large number of images that needed to be rated, this point was waived.

For one of the nonprofessional artworks, data is missing. This lack of ratings is due to the fact that only complete datasets were used and none of the experts, who rated this particular image finished the survey. However, it can be assumed that these values are missing at random and the sample size is still within sufficient size.

As with other art psychology research, this approach includes features of Eurocentrism and Americanism, in particular in three aspects: underlying art theory, available image material and

raters. First, the underlying theory on image analysis stems from European theorists and a particular way of thinking. Although these paradigms have been dominant in art history, they are by no means the only valid perspectives on art. Second, WikiArt (2021) provides a variety of international professional art, follows a knowledge equity policy and allows not only established institution to contribute, but also the public (Marengo et al., 2017). Nevertheless, with a marginalization of other artists, the overall ratio of images still overrepresents European and Northern American artists. Within the sample of nonprofessional artworks, we aimed for an international sample. But the call in English and German must be seen as a barrier to participation for people who speak neither nor. Third, the current questionnaire and survey were in German. Thus, the majority of experts stemmed from Germany, Austria or Switzerland.

Implications

As the first reliable, validated measurement RizbA allows researchers to not only capture psychological and physiological variables, but also the artwork itself. For example, which formal elements make viewers appreciate art? Is it representation, color mixture, or shaping? Is it easier for an audience distant from art to appreciate Van Gogh's works, rather than connecting to contemporary art by Firelei Báez? Since there are correlations between art preferences, thinking styles and personality (Gridley, 2006, 2013; Silvia & Nusbaum, 2011) these could be investigated in more depth. For example, an attitude of openness to experience positively correlates with a liking for abstract art (Gridley, 2013). But are individuals with higher scores on openness also more likely to create abstract art, when given the opportunity? More examples of how RizbA enhances art psychological methodology is extensively discussed in previous publications (e.g., Schoch et al., 2017; Schoch & Ostermann, 2020, 2021, May 4). However, besides offering a reliable measurement, no stable factor solution for pictorial expression could be found.

Bockemühl (1989) cites a tilt figure as an example for various perspectives on one and the same object, which cannot be perceived simultaneously. He states that aesthetics, if seen as a practical category, cannot be dealt with solely by logic. This might also be an analogy to analyzing art: It is highly ambiguous, refers to various levels and offers numerous perspectives of analysis. Perhaps a model simplification, as used in quantitative research alone is insufficient to capture artworks.

Future research

First of all, the theory-based model C should be tested on an even larger dataset of several thousands of images, in order to discover whether a stable solution can be found. Moreover, further postdisciplinary research is needed to evolve a larger theoretical model of image analysis that does justice to art. Finding substantial differences between theory and empirics would fuel the discourse between arts, humanities and science described by Mersch (2019) and other theorists. Only by bringing those perspectives together, might we one day be able to empirically and globally capture art.

A next step is validating the English translation of the questionnaire to provide a reliable scale. Other language versions could be translated and validated as well. This would lay the foundation to recruit raters and data from other regions and contexts in order to put the Eurocentric view into perspective. Further validation of other image samples should be pursued, such as non-handmade techniques, images created by children and adolescents and in particular by creators from all geographical regions and origins. Further studies could be multi-centered, collaborating with experts from different regions adding a critical view on Eurocentrism (Mosquera, 1992). Other approaches on art should be included in research on image analysis, such as from Africa, Asia, Oceania and Southern America, not in colonial continuities, but as equally valid points of view. Going further, those future research projects should not only

include academic theories on how to analyze art, but also knowledge, that has traditionally been excluded from previous research, such as emic indigenous perspectives of art.

Conclusion

A reliable intersubjective measurement of pictorial expression is possible using RizbA. However, no stable factor structure with an acceptable model fit throughout could be found. Comparing the models, model A and B yield higher internal consistencies than model C. The factors of model C correlate less with each other than those of model A and B. Nevertheless, model C is the only one with a converging model during a CFA that might provide a good model fit within a larger image sample. It may just as well be, that unlike inherent psychological variables, for the construction of pictorial expression there is no stable factor structure that is globally valid over all sorts and levels of art. Art is highly ambiguous, refers to various levels and offers numerous perspectives of analysis. Perhaps a model simplification, as we use it in quantitative empirics alone, is insufficient. Further post-disciplinary research is needed to evolve a larger theoretical model of image analysis that does justice to the subject. Only by this might we one day be able to globally capture art.

Studies show that RizbA is able to reliably measure pictorial expression. Going a step further this study revealed intriguing difference between empirics and art theory. Leder et al. (2012) might be right – at least for today’s perspective – when they stated that art is a mysterious aspect of human experience. In short: It’s complicated and yet a promising starting point for future research on art that should think postdisciplinarily.

References

- Apple. (2020). Photos for MacOS. <https://www.apple.com/macOS/photos>
- Arnheim, R. (2013). *Kunst und Sehen: eine Psychologie des schöpferischen Auges [Art and seeing: A psychology of the creative eye]*. Walter de Gruyter.
- Bauer, H. (1996). Form, Struktur, Stil: Die formanalytischen und formgeschichtlichen Methoden [Form, structure, style: The methods of form analysis and form history]. In H. Belting, H. Dilly, W. Kemp, W. Sauerländer, & M. Warnke (Eds.), *Kunstgeschichte: Eine Einführung [Art history: An introduction]* (Vol. 6, pp. 157-173). Reimer.
- Betensky, M. G. (1991). Kunsttherapie und künstlerische Äußerung aus phänomenologischer Sicht [Art therapy and artistic expression from a phenomenological perspective]. In J. A. Rubin (Ed.), *Richtungen und Ansätze*

- der Kunsttherapie: Theorie und Praxis [Approaches to art therapy: Theory and technique] (pp. 176-183). Gerardi.
- Bockemühl, M. (1989). Anschauen als Bildkonstitution [Looking as image constitution]. In C. Fruh, R. Rosenberg, & H.-P. Rosinski (Eds.), *Kunstgeschichte – aber wie? [Art history – but how?]* (pp. 63-82). Reimer.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155.
- Davis, L. (2018). WikiArt Retriever API. In <https://github.com/lucasdavid/wikiart>
- DeCoster, J. (1998). Overview of factor analysis. <http://www.stat-help.com/notes.html>
- DeepL. (2020). DeepL. In <https://www.deepl.com/translator>
- Eber, P., Müller, S., Bader, R., & Baukus, P. (1998). Die Nürtinger Beurteilungsskala zur Einschätzung von Patientenbildern im Therapieverlauf [The Nürtinger Rating Scale for evaluating pictures of patients in the therapeutic process]. *Fachhochschule für Kunsttherapie Nürtingen*.
- Eberlein, J. K. (2008). Inhalt und Gehalt: die ikonografisch-ikonologische Methode [Content and substance: The iconographic-iconological method]. In H. Belting (Ed.), *Kunstgeschichte: eine Einführung [Art history: An introduction]* (pp. 175-197). Reimer.
- Elbing, U., & Hacking, S. (2001). Nürtinger Beurteilungsskala und Diagnostic Assessment of Psychiatric Art: Neue Wege zur Evaluation der Bilder von Kunsttherapie-Patienten [Nürtinger Rating Scale and Diagnostic Assessment of Psychiatric Art: New ways of evaluating pictures of art therapy patients]. *Musik-, Tanz- und Kunsttherapie*, 12(3), 133-144. <https://doi.org/10.1026/0933-6885.12.3.133>
- Epskamp, S., Stuber, S., Nak, J., Veenman, M., & Jorgense, T. D. (2019). *semPlot version 1.1.2*. In <https://cran.r-project.org/web/packages/semPlot/index.html>
- Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on It. *European Journal of Psychological Assessment*, 33, 399-402. <https://doi.org/10.1027/1015-5759/a000460>
- Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., Walker, S., Winsemius, D., & Zeileis, A. (2019). *car version 3.0-10*. <https://cran.r-project.org/web/packages/car/index.html>
- Genz, A., & Kenkel, B. (2015). *pbivnorm version 0.6.0*. <https://cran.r-project.org/web/packages/pbivnorm/index.html>
- George, D., & Mallery, P. (2003). Reliability analysis. *SPSS for Windows, step by step: a simple guide and reference*, 14th edn. Boston: Allyn & Bacon, 222-232.
- Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research to Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus.
- 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.
- Hacking, S., Foreman, D., & Belcher, J. (1996). The descriptive assessment for psychiatric art: A new way of quantifying paintings by psychiatric patients. *The Journal of nervous and mental disease*, 185(7), 425-430. <https://doi.org/10.1097/00005053-199607000-00005>
- Hu, L. t., & 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. <https://doi.org/10.1080/10705519909540118>
- Huber, H. D. (2005). Systemische Bildwissenschaft [Systemic visual culture]. In K. Sachs-Hombach (Ed.), *Bildwissenschaft zwischen Reflexion und Anwendung [Visual culture between reflexion and application]*. Von Halem.
- 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]. Open Science Framework. <https://osf.io/nsw49>
- Kandinsky, W. (1955). *Punkt und Linie zu Fläche: Beitrag zur Analyse der malerischen Elemente [Point and Line to Plane: Contribution to the analysis of pictorial elements]*. Benteli.

- Kemp, W. (1988). Kunstwerk und Betrachter: Der rezeptionsästhetische Ansatz [Art work and viewer: The reception-aesthetic approach]. In *Kunstgeschichte: Eine Einführung* [Art history: An introduction] (pp. 241-258).
- 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>
- Leiner, D. (2018). *SoSci Survey*. <https://www.soscisurvey.de>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior research methods*, 48(3), 936-949.
- 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>
- 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* HCI International 2017 – Posters' Extended Abstracts. HCI 2017. Communications in Computer and Information Science, Cham.
- Marotzki, W., & Stöetzer, K. (2006). Die Geschichten hinter den Bildern: Annäherungen an eine Methode und Methodologie der Bildinterpretation in biografie- und bildungstheoretischer Absicht [The stories behind the pictures: Approaches to a method and methodology of image interpretation in biography and educational theory]. In M. Winfried & N. Horst (Eds.), *Bildinterpretation und Bildverstehen: Methodische Ansätze aus sozialwissenschaftlicher, kunst- und medienpädagogischer Perspektive* [Image interpretation and image understanding: Methodological approaches from a socioscientific, artistic and media educational perspective]. VS Verlag für Sozialwissenschaften.
- 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.
- Meyer, G. (2011). *Sprache der Bilder: Kunst verstehen: Form, Farbe, Komposition* [Language of the pictures: Understanding art: Shape, color, composition]. Seemann.
- Miyoshi, T. (2020). *Contact Form 7*. <https://contactform7.com>
- Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire design] (2 ed.). Springer.
- 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>
- Natterer, M., Pagès, J., Kolås, Ø., & Budig, S. (2020). *Gimp: GNU Image manipulation program*. <https://www.gimp.org>
- 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.
- Pacher-Nishikawa, A., Heck, T., & Schoch, K. (2021). Open editors: A dataset of scholarly journals' editorial board positions. SocArXiv <https://doi.org/10.31235/osf.io/jvzq7>
- Panofsky, E. (1978). *Sinn und Deutung in der bildenden Kunst* [Meaning in the visual arts]. DuMont.
- Panofsky, E. (2006). *Ikongraphie und Ikonologie: Bildinterpretation nach dem Dreistufenmodell* [Iconography and iconology: Image interpretation according to the three-stage model]. DuMont.
- Parry, S. (2017). Fit Statistics commonly reported for CFA and SEM. *Cornell Statistical Consulting Unit: Cornell University*, 2.
- Revelle, W. (2016). *psych version 2.0.12*. <https://cran.r-project.org/web/packages/psych/index.html>
- Rosseel, Y. (2020). *lavaan version 0.6-7*. <https://cran.r-project.org/web/packages/lavaan/index.html>
- Rosseel, Y. (2021). The lavaan tutorial. Ghent University. <https://lavaan.ugent.be/tutorial/tutorial.pdf>
- 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. <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. (2020). Giving the art greater weight in art psychology: RizbA, a quantitative questionnaire for formal picture analysis. *Creativity. Theories – Research – Applications*, 7(2), 373-410. <https://doi.org/10.2478/ctra-2020-0019>
- Schoch, K., & Ostermann, T. (2021, May 4). Psychometrics of art: Validation of RizbA, a quantitative rating instrument for pictorial expression. SocArXiv. <https://doi.org/https://doi.org/10.31235/osf.io/uc9e2>

- Sen, S., Acar, S., & Cetinkaya, C. (2014). Development of the Person–Environment Fit Scale (PEFSC): A new measure of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 8(4), 433. <https://doi.org/10.1037/a0037724>
- Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, 79(2), 310-334.
- 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/10.1037/a0021914>
- 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>
- Stuhler-Bauer, A., & Elbing, U. (2004). Kritik standardisierter empirischer Bilderfassung aus kunsttherapeutisch-phänomenologischer Sicht: Die Nürtinger Beurteilungsskala und die Diagnostic Assessment of Psychiatric Art, deutsche Version [Review on standardized empirical picture analysis from an art therapeutic phenomenological view: The Nürtinger Rating Scale and the Diagnostic Assessment of Psychiatric Art, German version]. *Musik-, Tanz- und Kunsttherapie*, 15(1), 5-15. <https://doi.org/10.1026/0933-6885.15.1.5>
- Suhr, D. D. (2005). Principal Component Analysis vs. Exploratory Factor Analysis. *Proceedings of the thirtieth annual SAS® users group international conference*, 203, 30. <http://www2.sas.com/proceedings/sugi30/203-30.pdf>
- Suhr, D. D. (2006). Exploratory or confirmatory factor analysis? *Statistics and Data Analysis*, 31. https://www.researchgate.net/publication/318722840_Exploratory_or_Confirmatory_Factor_Analysis
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10.
- Uzelac, M. (1998). Art and phenomenology in Edmund Husserl. *Axiomathes*, 9(1), 7-26.
- Vollmar, K. (2008). *Die faszinierende Welt der Farben: ein Glossar von A-Z [The fascinating world of colors: A glossary of A-Z]*. ars momentum.
- Walter, S., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in medicine*, 17(1), 101-110.
- Wickham, H., & Jennifer Bryan. (2013). *readxl version 1.3.1*. <https://cran.r-project.org/web/packages/readxl/index.html>
- Wiesing, L. (2005). Methoden der Bildwissenschaft [Methods of visual culture]. In K. Sachs-Hombach (Ed.), *Bildwissenschaft zwischen Reflexion und Anwendung [Visual culture between reflexion and application]*. Von Halem.
- WikiArt. (2021). *WikiArt: Visual art encyclopedia*. <https://www.wikiart.org/>
- Wordpress. (2020). *WordPress*. <https://wordpress.com>

Appendix

Acknowledgements

The authors gratefully acknowledge Thomas Gengenbach for providing theoretical and practical support regarding IT. Thanks also to Rebecca Kahn for the feedback on the manuscript.

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).

Abbreviations

API	Application programming interface
CFA	Confirmatory factor analysis
CFI	Comparative fit index
<i>df</i>	Degrees of freedom
GLAM	Galleries, libraries, archives, museums
MLR	Robust maximum likelihood estimator
PCA	Principal component analysis
PHP	PHP: Hypertext Preprocessor
RizbA	Ratinginstrument für zweidimensionale bildnerische Arbeiten (Rating instrument for two-dimensional pictorial works)
RMSEA	Root mean square error of approximation
SRMR	standardized root-mean-square residual
TLI	Tucker-Lewis Index

Additional Material

All material is available under a CC-BY 4.0 license as Open Data (image sample, metadata, ratings) and Open Methodology (SoSci Survey structure including PHP code for randomization and drawing, WordPress plugin Contact Form 7 code, R script) via Open Science Framework: https://osf.io/jnz67/?view_only=cd59189930484270aaa1f83017bee7e0.

A paper-pencil-version version of RizbA is available as Open Methodology under a CC-BY 4.0 license:

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].

Zenodo. [doi:10.5281/zenodo.3765221](https://doi.org/10.5281/zenodo.3765221)

Table 9

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