Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 44 (2023) 48–50

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Comment

Can we “read” art... at all? A psychometric perspective on the possibility of measuring artwork attributes Comment on “Can we really ‘read’ art to see the changing brain? A review and empirical assessment of clinical case reports and published artworks for systematic evidence of quality and style changes linked to damage or neurodegenerative disease” by Pelowski et al. (2022)

Nils Myszkowski

Department of Psychology, Pace University, United States

Received 28 November 2022; accepted 29 November 2022

Communicated by J. Fontanari

Pelowski et al. [12] make an exciting proposition: “reading” art to infer on their author’s neuropsychological condition – present or past. By “reading” art, they more specifically suggest using some form of measurement of aesthetic and creative features of artistic products. The proposition is remarkable from a substantive point of view. Further, the results of their research indicate that evaluations of art may, indeed, have some predictive validity over certain conditions. As the authors acknowledge, this should be considered a first effort to put together existing research on the topic and to provide a systematic reanalysis of the artwork used in this research. Putting together and making sense of such scattered literature, while building a bridge between two distant domains, is certainly a *tour de force* worth applauding. It appears to me, nevertheless, that attributes of artworks – especially aesthetic quality and creativity – are often thought as inherently subjective (“in the eye of the beholder”). Thus, it is perhaps shocking to consider characteristics of artworks as possible objects of measurement (e.g., [6]), let alone to think of them as symptoms of medical conditions and events. Indeed, the idealization of Van Gogh’s genius might be somewhat spoiled by thinking of it as a symptom of temporal lobe dysfunction. Although I am not competent to speak for the neurological aspects, in this commentary, I want to discuss an important assumption of the authors’ research: That it is possible to “read” art scientifically in the first place. I will briefly argue that (1) all artistic features can be thought as suitable for measurement (even artistic quality) and that (2) modern test theory is an ideal framework to control for subjectivity (i.e., rater effects) in art measurement. I will finally (3) advance that artificial intelligence ratings also induce rater effects, and therefore do not solve the problem of objective measurement.

Are artwork attributes suitable for measurement? Of course, the measurement of very specific features like color depth is rarely questioned, but it is often advanced that other attributes (especially aesthetic quality) are not measurable

DOI of original article: https://doi.org/10.1016/j.plrev.2022.07.005.
E-mail address: nmyszkowski@pace.edu.

https://doi.org/10.1016/j.plrev.2022.11.010

(e.g., [6]). I would argue that behind this view is a moral statement ("the mystique of art must be sheltered from scientific inquiry") and/or an admission of failure ("it is impossible because we have not succeeded"). The former argument may be debated philosophically, but, deciding that some domains should remain unstudied is probably unscientific, as it precludes the pursuit of knowledge. Besides, how to decide between what can be studied and what is "too inscrutable to study"? By that logic, perhaps nature itself is too inscrutable. The second argument is a fallacy: failing to measure an attribute (well), even repeatedly, does not make a construct unmeasurable. In other words, "bad measures do not imply bad constructs" ([9], p. 660). Moreover, a number of studies have actually suggested that individuals spontaneously agree on attributes of artworks, even aesthetic quality (e.g., [7,10]). Classical and modern test theory inform us that, in measurement, observed behaviors (e.g., rater 1's judgment of artwork 1 as of "high quality") are thought as caused by a latent attribute (e.g., artwork 1's quality) – with some effects induced by the measurement device (typically, item and/or rater effects). This is often referred to as reflective measurement, or the causal theory of measurement [3], and this causal relation has been consequently proposed as the main requisite for valid measurement [4]. In our context, it is trivial that the quality of an artwork can be realistically thought as a cause for individuals judging it as of quality (otherwise, there would be no museums to display certain artworks while others are not displayed). Consequently, I argue that *all* features of artworks (even their overall quality, or their beauty) can – at least, in theory – be measured. Pelowski et al. [12] demonstrate an important reason why they should.

But how to do it well? The authors use up-to-date methods, mainly relying on human rater judgments, a classical paradigm in the field [1]. Nevertheless, their work suffers from the scatteredness and inconsistency of the recommendations on rater-mediated assessment in creativity research [5,11]. Notably, their methods, based on linear mixed models (with intraclass correlation coefficients) make unrealistic assumptions regarding their data. The main ones are that the latent attribute is linearly related to the rating, that the ratings are normally distributed (they are actually ordinal here), and that all judges are equally discriminant. While I do not advance that the validity of their findings is jeopardized here, it is clear that our field has yet to develop a consistent and robust methodological approach for rater-mediated measurement. As I have argued [8,11], item-response theory (IRT) provides a unified framework that makes realistic assumptions and allows to account for item and rater effects (varying severities, difficulties, discrimination, etc.), along with their interactions [11,13,14]. Further, seeing how dispersed the literature is on reliability estimation with human raters in our field [5], and how it is usually disjointed from scoring procedures (e.g., the same average score is used regardless of the reliability estimate) [11], having a framework that uses a single model to estimate attributes (e.g., to measure an artwork's quality) and their uncertainty (i.e., standard error/reliability) is a key advance. Since research on artistic attributes often uses rater-mediated assessment, accounting for rater subjectivity is central to accurate measurement. IRT elegantly addresses this problem and should be more widely used if we are to take human ratings of art seriously in research.

Finally, I want to make a note regarding the use of computer judgments and their reliability – since they are also used by the authors, are an emerging tool (e.g., [2]), and since the authors extensively justify using human ratings, in spite of subjectivity issues (p. 81). The fact that a computer algorithm can be programmed to provide reproducible ratings is certainly significant. However, reproducibility does not imply objectivity. In other words, one should not lose sight of the fact that other algorithms could have been used and could have provided different results, in the same way that other human judges could have been used. In other words, there is some uncertainty inherent to using judges, human or not. Consequently, computer rater-mediated measurements need to be psychometrically inquired just as much as human rater-mediated measurements. In other words, computerized ratings do not solve the problem of objective measurement: instead, they replace human subjectivity with synthetic subjectivity. In the current state of things, I believe the authors made an excellent call in using both.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Amabile TM. Social psychology of creativity: a consensual assessment technique. J Pers Soc Psychol 1982;43(5):997–1013. https://doi.org/10.1037/0022-3514.43.5.997.

[2] Beaty RE, Johnson DR. Automating creativity assessment with SemDis: an open platform for computing semantic distance. Behav Res Methods 2020. https://doi.org/10.3758/s13428-020-01453-w.

[3] Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev 2003;110(2):203–19. https://doi.org/10.1037/0033-295X.110.2.203.

[4] Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. Psychol Rev 2004;111(4):1061–71. https://doi.org/10.1037/0033-295X.111.4.1061.

[5] Cseh GM, Jeffries KK. A scattered CAT: a critical evaluation of the consensual assessment technique for creativity research. Psychol Aesthet Creat Arts 2019;13(2):159–66. https://doi.org/10.1037/aca0000220.

[6] Gear J. Eysenck's Visual Aesthetic Sensitivity Test (VAST) as an example of the need for explicitness and awareness of context in empirical aesthetics. Poetics 1986;15(4–6):555–64. https://doi.org/10.1016/0304-422X(86)90011-2.

[7] Mitrovic A, Hegelmaier LM, Leder H, Pelowski M. Does beauty capture the eye, even if it's not (overtly) adaptive? A comparative eye-tracking study of spontaneous attention and visual preference with VAST abstract art. Acta Psychol 2020;209:103133. https://doi.org/10.1016/j.actpsy.2020.103133.

[8] Myszkowski N. Development of the R library "jrt": automated item response theory procedures for judgment data and their application with the consensual assessment technique. Psychol Aesthet Creat Arts 2021;15(3):426–38. https://doi.org/10.1037/aca0000287.

[9] Myszkowski N, Çelik P, Storme M. Commentary on Corradi et al.'s (2019) new conception of aesthetic sensitivity: is the ability conception dead?. Br J Psychol 2020;111(4):659–62. https://doi.org/10.1111/bjop.12440.

[10] Myszkowski N, Storme M. Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). Pers Individ Differ 2017;117:91–100. https://doi.org/10.1016/j.paid.2017.05.041.

[11] Myszkowski N, Storme M. Judge response theory? A call to upgrade our psychometrical account of creativity judgments. Psychol Aesthet Creat Arts 2019;13(2):167–75. https://doi.org/10.1037/aca0000225.

[12] Pelowski M, Spee BTM, Arato J, Dörflinger F, Ishizu T, Richard A. Can we really 'read' art to see the changing brain? A review and empirical assessment of clinical case reports and published artworks for systematic evidence of quality and style changes linked to damage or neurodegenerative disease. Phys Life Rev 2022;43:32–95. https://doi.org/10.1016/j.plrev.2022.07.005 [this issue].

[13] Primi R, Silvia PJ, Jauk E, Benedek M. Applying many-facet Rasch modeling in the assessment of creativity. Psychol Aesthet Creat Arts 2019;13(2):176–86. https://doi.org/10.1037/aca0000230.

[14] Robitzsch A, Steinfeld J. Item response models for human ratings: overview, estimation methods, and implementation in R. Psychol Test Assess Model 2018;60(1):101–39.