# Psychometric Equating Methods Ease Fitness Function Dilemma in Generative Art

**Dr. Nikolaus Bezruczko, PhD**
*Department of Clinical Psychology*
*Chicago School of Professional Psychology*
*Chicago, USA*
nbezruczko@msn.com

## Premise

Generative art is responsible for many new ideas in visual arts and music and has contributed to understanding fundamental mechanisms affecting human development and behavior. Practical accomplishments making art have also pointed to new ideas about creativity and aesthetics. Yet, despite these advances generative art still lacks capacity to evaluate its own artistic products. Typically referred to as "fitness bottleneck", autonomy and expression associated with automated generative art is presently constrained by an incapacity to identify its best work or objectively compare independent image sets. This report presents a psychometric strategy on this issue, as well as examples demonstrating how those constraints could be eased by integrating psychometric aesthetic derived from professional artist judgments into fitness evaluation. Parameterized standards and adaptation of psychometric common item equating methods could guide generative productions and clarify their aesthetic value without reoccurring costs commonly associated with expert jury panels. Results presented here show a diverse group of professional artists converge on an objective aesthetic standard that could be integrated into generative algorithms. Successful empirical studies suggest this goal is reasonable.

## 1.0 Introduction

Michael Noll's surprising demonstration of a computer generating Mondrian's Composition with Lines indistinguishable from original work remains a landmark in generative art [1]. This capacity to reproduce authentic art, while not yet autonomous, continues in contemporary generative arts through elaborate decomposition and assembly methods that mimic artists such as Kandinsky, Miro', and Pollock [3]. Generative visual arts are now being expressed broadly in genetic algorithms and programming, evolutionary arts, computational aesthetics, as well as seemingly endless emerging forms of autonomous visual arts. Several approaches to image productions implementing algorithms and rule-based systems now easily provide efficient and frequently unusual, if not creative, images without artist intervention. Some commentators have, arguably, asserted that generative art will extend human creative capacity [2], while they have already expanded methods of expression.

Yet, despite extraordinary advances producing unique, autonomous, computer generated art, an annoying problem is evaluation of image quality, which typically remains dependent on expensive and time-consuming juries [4]. Without an appropriate aesthetic value function that efficiently evaluates images, evolutionary art process is dramatically slower. This limitation presents a serious practical challenge applying EC approaches to fields such as image and music generation. Todd and Warner [5] referred to this inability to measure aesthetic quality simultaneous with image production, the fitness bottleneck.

Many fitness bottleneck strategies have attempted to alleviate practical implications of manual intervention during image production. Passive and interactive methods have been proposed that either guide image replications during image production toward expected aesthetic values or impose selection standards after image runs. While interactive methods are typically implemented with an artist, even automated approaches in contemporary evolutionary art require artists somewhere in the image production loop manually scoring each new image or more specifically phenotype. Consequently, this limitation logically undermines any conception of full automation, threatening idea of genuinely autonomous art if not eliminating it.

According to some authorities, a more profound aspect of this problem is fully automated fitness evaluation is, in principle, conceptually inconsistent with contextual foundations of contemporary aesthetic theories, which embed emergent art in social systems. According to this perspective, valid fitness functions for evolutionary art systems cannot exist or be justified independently of their social context., which have never been addressed in generative arts. How does an algorithm embody cultural properties?

A further complication are dyspeptic convulsions that arise between modern and postmodern attitudes toward aesthetic fitness evaluation. Especially image selection that imposes an arbitrary algorithm stopping value intended to emulate an objective aesthetic standard. Cost and convenience overriding cultural relativism creates tension between modern beliefs about objective standards and postmodern commitments. This conflict concerning aesthetic outcomes leads to attacks on legitimacy of sorting images into quality categories.

Even technical procedures seem to present intractable issues such as inferential implications of typical evaluation juries. For example, a common method is arbitrary sample-based standards defined by consensual agreement among jury members, which is subject to sampling variability. Aesthetic evaluation, even when juries are constituted by artists and expert judges, are inherently unstable, which weaken validity of image orders, juries, samples, and algorithms. Not surprisingly, mounting difficulties of practical fitness evaluation seem to make any hope of improving image selection seem futile if not hopeless.

Purpose of this report is to offer an alternative to conventional fitness evaluation methods by describing an adaptation of psychometric item scaling methods widely applied in rehabilitation medicine outcome evaluation, educational and psychological measurement, as well as licensure and certification examinations to generative arts. Psychometric scaling methods are implemented to statistically equate generated test items across forms and item banks, as well as population samples. While physical units differ -- visual arts images versus mental test items -- concerns about qualitative evaluation, as well as stability of quantitative invariance across item and image pools are comparable. In other words, issues commonly involved with authentic aesthetic evaluation are fundamentally like those addressed in mental test development.

Psychometric equating methods in psychology and education present an interesting perspective on problem of evaluating generative art. Practical implementation of mental test theory is highly dependent on objective standards, and instrumentation is typically conducted with generated item samples, which themselves may have been produced randomly from an algorithmic process, and item replications are typically accompanied by qualitative variation. Hence an item scaling method that equates test forms to objective standards is essential to maintain item and form comparability across population samples. Adaptation of this methodology to evolutionary art images could alleviate variability presently associated with disparate images from idiosyncratic algorithms, as well as identify objective aesthetic value.

In this research, a pool of generated visual images was first evaluated by a professional artist jury with Likert rating scales to describe approximate aesthetic quality, which provided numerical values for an ordinal image ranking. Mathematical transformation of ratings with a logistic function constructed a framework where image values, as well as algorithm specification codes were parameterized as logits, which function as objective weights. This parameterized image ranking then provided foundations for an aesthetic dimension with statistically invariant properties and estimates of standard errors and psychometric reliability.

Images scaled to this framework then were useful for statistical equating of future image generations, as well as images generated by modified algorithms but without jury implementation. Seeding procedures in this context were also conducted. An application is presented here both of automatic item generation and image “seeding” during interactive implementation with an artist.

While this methodology does not eliminate expert panels or professional artist judgments, which in principle is impossible, this demonstration provides an objective method for comparing items from different production runs, under certain conditions from different algorithms, as well as different media expressions without conducting additional juries. This strategy involves an initial investment in validity that is recovered by diminishing burden of reoccurring fitness evaluations.

This report consolidates advances in psychometric scaling and equating methods with insights from empirical aesthetics but with explicit emphasis on professional artist validation. Many computational aesthetic studies have attempted to integrate fitness evaluation and empirical aesthetics into image production but without convergence of professional artists. Those approaches are irrevocably inconsistent and distinct from methods presented here.

## 2.0 Background

Image fitness in generative arts commonly refers to aesthetic quality of images yielded by an algorithm intended for an explicit purpose or target audience. Ideally, fitness would be established by a sensitive discriminative function that evaluates image quality from generated population and classifies aesthetic acceptability. Consequently, fitness functions establish aesthetic boundaries for accepting works produced from algorithms, and their central goal is to clarify correspondence between generated images and desired level of aesthetic quality.

Practical limitations of fully automated algorithmic aesthetic evaluation have led to interactive methods that require intervention by expert observers who assign scores or ratings. In general, interactive evaluations do not recruit professional artist samples to justify parameters for aesthetic evaluation. Instead, empirical rank orders are based on convenience samples though sometimes very large.

Typical fitness strategies do not address or solve and, arguably, cannot solve fundamental issue of aesthetic standards, which arbitrarily fluctuate among sample-based evaluation juries, and their validity is further complicated by cultural context. Nonetheless, efforts to systematize fitness evaluation has moved ahead aggressively along several approaches with varied success.

Three prominent strategies to fitness evaluation are:

- Interactive models
- Evolving genotypes
- Arbitrary aesthetic measures
- Corpus methods

Interactive judgment models are least desirable but remain dominant. Evolving genotypes are automated models that rely on internal and/or external standards, which force generated images to converge on declared aesthetic standards. Seeds and targets also impose external standards on image evolution, then algorithms run their course. An alternative automated system may implement dimensional extraction models (principal components), which are imposed on obtained phenotypes [6]. Finally, corpus methods implement deep learning networks, which identify underlying properties across immense data bases of successful art works and parameterize their emergence in generated art.

## 2.1 Computational aesthetics approaches to evaluation

### 2.1.1 Automated fitness functions

A goal of computational aesthetics and evolutionary arts is to fully automate fitness evaluation simultaneously with image production, which would guide multiple iterations to convergence on optimal aesthetic values. Their goal is to address limitations and constraints of manual models. Machado, Romero, and Manaris identified “essentially five approaches to fitness assignment: interactive evolution, similarity based – evolving towards a specific image or images, hardwired fitness functions, machine-learning approaches, and co-evolutionary approaches (p. 383) [7]”.

Physical properties have been examined for their contribution to image quality. For example, Heijer and Eiben [8] compared four methods of measuring fitness: processing complexity model based on image compression ratio, Ralph’s bell curve, fractal dimension peak, and their aggregated or weighted sum. Unfortunately, they found little agreement among them. In addition to fractal dimension [9], fitness evaluation has also been based on physical image characteristics such as GIF compression [10], overall luminance gradient strength [11], or edge density [12]. Likewise, color and contrast belong to low-level image properties that can affect the preference ratings of photographs. Authors argue that artists use a non-linear compression to obtain low skewness in their paintings because images with this property can be more efficiently processed by the visual system. Inconsistency of above measures have led researchers to explore insights from psychological studies of aesthetic preference, disparate as they may be.

Advances in computing hardware and methodology have accelerated attempts to integrate affective image properties based on psychological empirical studies that are known to influence human preference with physical image properties described above. This approach is expected to improve both automated and adaptive approaches to aesthetic image evaluation [13, 14]. This general effort to integrate empirical aesthetics into generative art production is now called computational aesthetics (CA). Unfortunately, literature produced by psychological empirical aesthetic studies is vast and inconsistent, arguably based on weak methods, which present substantial challenges to understanding implications for human judgements of beauty or contribution to aesthetic experience.

> Computational aesthetics" is sometimes used in the sense of describing a class of artefacts made by computers . . . we will refer to computational aesthetics only as computational models of human aesthetics [14].

### 2.1.2 Deep learning neural networks

An alternative approach to integrating physical and affective image properties in fitness models is automatic feature learning, which implements deep neural network methods. Central goal here is to incorporate heterogeneous inputs generated from images of authentic visual art from both global and local perspectives, then unify extracted information into a predictive model. Applications have been presented with AVA dataset [15]. A related strategy is decomposition or separation of image style and content using convolutional neural networks [16]. See also brain inspired deep networks [17].

Deep learning networks use authentic art images as a training set to identify common properties that mimic authentic images. "Even in their perceived autonomy as image creators, their ability to act autonomously is limited within a very tight statistical framework that is derived from their training data [18]." Consequently, generative aspect of this system is constrained by the training set. This issue is echoed by other researchers as well. "AI systems that are trained to extract features from curated data-sets constructed of contents produced by people are imitating properties of artefacts rather than autonomously searching for novel means of expression. This holds true for current, popular AI art systems using machine learning [18]."

Extraction of fitness models using deep learning methods forces generative arts image production to conform to an aesthetic standard that is compatible with those images in the extraction pool. When large enough, those models can claim validity but raise questions whether these powerful systems are sacrificing autonomy for expediency. In other words, implementation of deep learning algorithms creates fitness dependency on the extracted learning and imposed on the generating function [18].

### 2.2 Challenges associated with sample-based, unstandardized, and non-validated fitness models

Traditional approaches to evolutionary art fitness evaluation have relied on sometimes naïve, expedient solutions emphasizing procedural convergence. Current trend emphasizes more understanding of empirical studies of aesthetic preferences and implications for fitness evaluation [13, 14]. Yet, Lewis [19] described numerous challenges to automating fitness evaluation in visual arts (see pp. 24-26), while Johnson [14] emphasized problems presented by differences in individual preference for image properties [20]. Not least of these challenges is long standing confusion concerning individual preference differences for affective image properties versus formal aspects of authentic artworks, which are discussed below.

### 2.2.1 Confusion related to psychological empirical aesthetics

While psychological studies of empirical aesthetics have increased awareness of objective image properties, which increases comprehensiveness of CA approaches to fitness, they also introduce enormous confusion. Many constructs in psychological aesthetic studies such as complexity, order, symmetry, and randomness are formulated so poorly that general trends are difficult to establish. In addition, multiple approaches to operational definitions have led to inconsistent results, and general trend of 20th century empirical aesthetics is lack of consensus about chief findings. Replications are typically sparse hence objective image properties are only partially understood, and CA implementation of them have been fraught with complications. Strongest ideas coming from empirical aesthetics with useful implications are related to complexity, uniformity, symmetry, and order, as well as rule of two thirds, but they are also among most notoriously inconsistent and misused.

Complexity, for example, as objective property has been studied extensively, both in generative arts and in psychological experiments in multitude of statistical measures. Most prominent is number of visual elements in an image [21], an objective frequency of image elements, which is central to information processing models. This measure is prominently correlated with several computerized models of complexity (Zimmer's Law) [22]. Zhang [2] examined visual complexity in psychological studies versus computational complexity. Yet, complexity role in visual art differs across common preference measures and becomes incomprehensible when integrated among sometimes vast arrays of more complicated image properties such as semantics and affective expression. Consequently, complexity is shrouded in mystery, and Martindale [23] has pointed to Berlyne's studies, which have been reported widely, but continue to befuddle researchers [24, 25]. Central issues limiting their usefulness were his method of replication and interpretation, as well as restricted population sampling.

Nonetheless, CA researchers show growing appreciation for importance of complexity and order on individual preference differences [20). Güçlütürk et al. discussed role of individual differences in clarifying function of complexity in aesthetic evaluation and emphasized need to study them further for contributions to CA, while other research have found preference for complexity and order of professional artists, as well as those identified with high visual arts aptitude to differ significantly from laypersons and non-artists [20].

CA researchers have discovered alternative measures of complexity such as image compression [26] for measuring complexity. Friedenberg and Liby [27] also discussed alternative complexity estimates such as density, number of blocks, GIF compression rate and edge length associated with perception of beauty of semirandom two-dimensional patterns. Agreement among them, however, is inconsistent, and validation studies are not typically conducted.

> These mixed results originate in variety of metrics used to estimate what is loosely called "complexity" in psychology and indeed refers to conflicting notions. We conclude that participants tend to prefer some types of complexity, but not all. These findings may help explain divergent results in the study of perceived beauty and complexity and illustrate the need

> to specify the notion of complexity used in psychology [28**].**

This confusion is not limited to complexity as questions and issues also surround symmetry, order, and quantifying symmetrical complexity [29, 30]. Complicating these matters further, stochastic variability or inherent randomness of an artwork is an important determinant of aesthetic preference [31] yet is not typically included in evolutionary algorithms.

### 2.2.2 Aesthetic validity of fitness evaluations

Among weakest aspects of fitness evaluation but receiving no attention is validation. Virtually all traditional fitness function approaches suffer from consensual scoring, which bases aesthetic standards on sample norms, however they may be defined. This approach first presented by Eysenck in 1940s is fraught with complications as sample variability ensures unstable aesthetic standards. This issue is compounded by failure to demonstrate professional artist convergence among samples. In general, undefined consensual standards are associated with unstable aesthetic standards, limited generalizability of fitness evaluations, and, ultimately, confusion about image fitness.

Other validity issues arise for studies that implement broad arrays of image features and properties without justification yet assert their function in judging aesthetic value. For example, Li and Chen examined 40 image properties and aesthetic preference without explanation of their functional role [32].

> Their connection to human aesthetic judgement is not clearly explained prior to their being employed as fitness functions. In some cases, the functions are called "measures" [18].

Brachmann and Redies [13] describe other validity issues associated with large website datasets in CA aesthetic investigations that use vast amounts of information both about images and samples. These images are typically evaluated online in an uncontrolled environment, which limits understanding their cultural context or professional composition of juries. Not surprisingly, virtually nothing is known about potential biases related to image author, popularity, reinforcement, display environment, and so on.

Cultural issues arise even when Image characteristics are implemented that have demonstrated empirical effectiveness predicting aesthetic preference such as randomness, complexity, and order. They could be instrumental during fitness evaluation, but isolated applications are problematic. In general, computational aesthetics recognizes limitations of generating images in isolation of cultural preferences [33]. Fortunately, importance of integration with artists is becoming recognized.

> Cultural contextual foundations limitations, which is related to Isolation of computational artists. In order to provide a cultural context to our agents, we propose their integration in a Hybrid Society of artists and critics, both computational and human [33].

While Lewis [19] went even further by emphasizing conflict when automated fitness algorithms are implemented without convergence of artists.

> Ultimately, how should results of automated fitness algorithms for evolutionary art be evaluated in a mixed culture of artists and computer scientists? Given two bodies of artistic images created using evolution, if knowledgeable computer scientists and computer

artists disagree about which ones are a success and which ones are a failure, what are the mechanisms by which research proceeds? What are the criteria by which progress can be evaluated [19]?

## 3.0 Psychometrics of aesthetic value

### 3.1 Historical background

Empirical measurement of aesthetic preference first appears in 19th century empirical studies Fechner conducted at Leipzig [34], which demonstrated feasibility of investigating relations between human perceptual judgments and physical dimensions. Those studies are now categorically referred to as psychophysics. Thurstone adapted Fechner's objective scaling methods in 1920s to measure human attitudes and mental abilities and suggested aesthetic applications [35, 36, 37]. Birkhoff [21] then provided mathematical foundations for an aesthetic measure, while Eysenck conducted empirical studies. Theoretical foundations for contemporary CA integration of empirical aesthetics is largely based on this body of research.

Those early quantitative methods became 20th century foundations for American mental testing movement, which elaborated and institutionalized mental measurement in education and psychology through the College Board and American Psychological Association. Equating methods presented in this report were initially part of that movement, and they were developed to systematize and justify evaluation of generated test forms across samples, as well as establish comparability of item banks. Equating methods have direct implications for establishing professionally validated fitness measures for generative visual arts. This historical development is briefly elaborated below.

#### 3.1.1 Fechner measures aesthetic preference

Contemporary approaches to empirical aesthetics are largely an extension of Fechner's seminal 19th century psychophysics research [34]. He empirically demonstrated systematic quantitative relations between perceptual responses and variation of physical stimulation, a landmark achievement that encouraged empirical investigation of human psychological experience. Prior to Fechner, Kant had concluded human experience existed beyond methods of science and rejected possibility of psychology ever becoming an empirical science.

Fechner's psychophysical methods involved study participants comparing physical specimen and rank ordering their weight differences, and he portrayed correspondence between perceptual judgments and physical stimuli with a mathematical function. Then he demonstrated group values conformed to a normal distribution with reproducible properties of mean and variance, and he asserted their standard deviation represents a just noticeable difference (JND). His approach inspired interest in experimental psychology and virtually all contemporary scaling methods are derived from his original insights.

#### 3.1.2 Early 20th century

Thurstone would establish important measurement foundations in 1920s when he developed objective methods for scaling test items [35, 36, 37] adapting Fechner`s methods measuring perceptual judgments. Thurstone in turn used JND to establish a scaling unit for measuring attitudes and opinions, then later mental abilities and achievement. His scaling structure would be elaborated to judgments about aesthetic stimuli. In 1933, Birkhoff [21] presented a mathematical formulation for aesthetic value based on proportional relation between complexity and order. Aesthetic measure increases, as complexity increases relative to order. He proposed two image characteristics, complexity and order, are functionally related to visual preference in following model:

Aesthetic measure = O/C

where M is an artistic measure that is a function of order and complexity. Practical interpretation is artistic value of any image is always greatest when order is maximized relative to complexity. At any level of complexity, an increase in order will increase overall aesthetic value.

Unlike many early scientific assertions, additional studies have corroborated this relation between complexity and order, which remains central in contemporary aesthetic theories. However, complexity measurement has become a contentious issue and social researchers have unsuccessfully sought alternative operational definitions that do not rely on Birkoff's insights of information density.

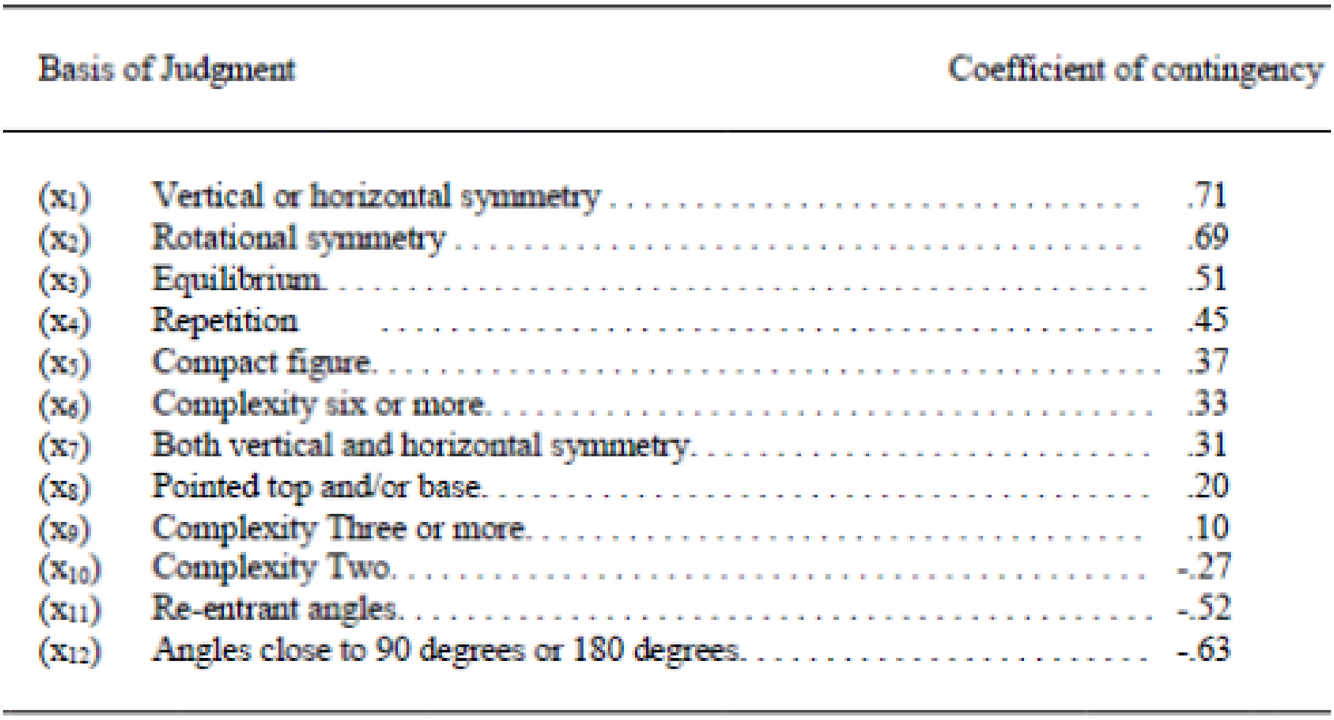
### 3.1.3 Eysenck

Eysenck extended Birkhoff's approach to aesthetic measure in several ways [38, 39, 40]. First, he factor-analyzed preference judgments for polygons and identified prominent type factors, which led to recognition of individual preference differences. T- and K-factors have received attention in subsequent empirical studies.

Eysenck's aesthetic measure formula from his published table below is (see Figure 1):

$$M = 20X_1 + 24X_2 + 8X_3 + \mathit{7X_4} + 5X_5 + 3X_6 + 3X_7 + 2X_8 + IX_9 - 2X_{10} - 8X_{11} – 15X_{12}.$$

Eysenck's general aesthetic factor or T-factor was eventually formulated into Visual Aesthetic Sensitivity Test (VAST). Unfortunately, VAST suffered from unreliability, then later failed validation by professional artists [41]. Contemporary efforts to resuscitate VAST have been largely unsuccessful [42]. Among Eysenck's discovery of aesthetic types, he found support for aesthetic judgment aptitude, a culturally transcendent construct.



| Basis of Judgment | | Coefficient of contingency |
|---|---|---|
| $(x_1)$ | Vertical or horizontal symmetry | .71 |
| $(x_2)$ | Rotational symmetry | .69 |
| $(x_3)$ | Equilibrium | .51 |
| $(x_4)$ | Repetition | .45 |
| $(x_5)$ | Compact figure | .37 |
| $(x_6)$ | Complexity six or more | .33 |
| $(x_7)$ | Both vertical and horizontal symmetry | .31 |
| $(x_8)$ | Pointed top and/or base | .20 |
| $(x_9)$ | Complexity Three or more | .10 |
| $(x_{10})$ | Complexity Two | -.27 |
| $(x_{11})$ | Re-entrant angles | -.52 |
| $(x_{12})$ | Angles close to 90 degrees or 180 degrees | -.63 |

Figure 1. Reproduced from Eysenck, 1941, Table 1, p. 89 [39].

### 3.2 Contemporary aesthetic measurement advances

#### 3.2.1 Information processing approaches of Shannon and Moles

Psychologists broadly agree visual stimuli consist of varied information sources that influence visual preference in some cumulative manner. Moreover, viewers are believed to extract information during image scanning and relay it to specialized neuron receptors where neurological processing reassembles a meaningful gestalt or percept. A key mechanism in this process is image decomposition during perception. Information theory is based on several principles of extraction 0riginally proposed by Shannon [43]. Uncertainty, amount of information, as well as information transmission and organization govern information in visual images, which have found broad commercial and scientific implications.

In the simplest application to aesthetics, Platt [44] proposed hierarchically organizing aesthetic information into discrete formal and stylistic levels, while Moles [45] proposed a more complex system that separates formal structural components of an image from independent semantic components that simultaneously superimpose meaning on comprehension during perception. He emphasized that "each level conveys its own unique message and possesses specific rules of organization [45]." His primary interest in semantic information represents systematic influence imposed by socially constructed entities such as religious, governmental, and educational institutions on visual art. Finally, Berlyne proposed expressive and syntactic levels [46, 47]. Expressive level transmits some personal aspect of artist, while syntactic information is physical configuration of visual elements in an object or pattern.

Information theory developments contributed considerable insight into understanding impact of aesthetic images on visual preference. Identification of discrete components and their systematic processing has also led to insights about hierarchical knowledge systems and facilitated understanding of efficient methods for transmitting visual information. Developmental studies have pointed to human individual differences that mediate responses to aesthetic information.

#### 3.2.2 Contribution of statistical models to aesthetic measurement

By mid-20th century, significant advances in statistical methods led to nonphysical (social) measurement models that eliminated troubling deficiencies of ordinal scores for quantifying human perceptual responses, which included attitudes, preferences, and opinions. In general, ordinal methods lack equal interval scale structures, and unit of measurement is unknown. Prominent among these advances were Rasch models [48], a probabilistic approach from logistic regression for transforming scores and ratings into objective, linear measures. However, unlike logistic regression, Rasch model estimation is not dependent on populations or specific samples, hence they provide quasi-absolute measures. Then Fischer introduced Linear Logistic Test Models [49], which permitted investigating hierarchal cognitive components underlying responses to items [49].

Statistical measurement models provided analytical frameworks for stochastic factors underlying preference for aesthetic images [31] and together with information theory led to test models such as Visual Designs Test [50] that became highly effective for visual arts aptitude evaluation. Validity of these advances for aesthetic theory only became clear with investigations of larger sample professional artist samples [51].

#### 3.2.3 Common item equating

Paralleling 20th century aesthetic measurement advances, large scale mental ability and college placement testing were being developed by American College Board. Practical needs to demonstrate comparability of multiple test forms with many multiple-choice test items led to scale equating methods [48, 52]. In addition, frequent testing required generating constantly new test items but maintaining qualitative properties such as difficulty across examination forms.

Today, common item equating is implemented world-wide for educational measurement, as well as professional licensure and certification to link item pools and disparate samples onto a common real number scale. Rasch models are widely popular for transforming ordinal item difficulty and person ability values to a common scale, and they have desirable properties of sufficiency, separability, simultaneous conjoint additivity, and specific objectivity. In addition, Rasch models are ontologically strong statistical models, which contribute to valid statistical invariance.

This report demonstrates that psychometric scaling methods can be adapted to evaluate aesthetic quality of visual images and are directly relevant to fitness evaluation in generative art. Figure 2 presents basic paradigm for item equating showing unique forms (image runs) being linked to an overarching dimension of common items.

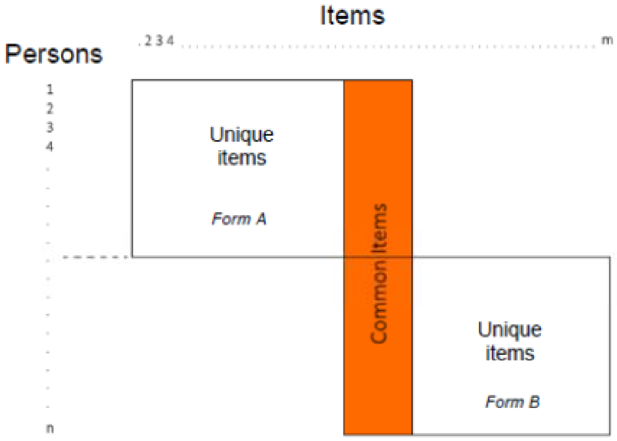


Figure 2. Schematic of horizontal common item equating procedure [48].

In general, items are first parameterized to identify locations on a linear scale, which establishes an explicit unit of measure. A subset of items is identified on this scale structure, and their aggregation functions as scaling constant for linking any other item pool sharing this subset to this structure. Ideally, linking items are embedded across qualitative content of evaluation dimension. See published applications of equating methods to automatically generated figural test items [53, 54, 55].

### 3.2.4 Adaptation of psychometric scale equating to generative art

Selection of optimal images during evolutionary art production is a challenge to contemporary generative arts. Consequently, an equating method validated by professional artists, which articulated an objective standard would be highly useful for identifying image fitness, as well as maintaining image quality across samples and algorithms.

In this research, images were submitted to a large sample of professional artists for judgment, and their ordinal scores or ratings, as well as image code specification were transformed to interval logits. After linear parameterization, this artistically validated dimension is an absolute aesthetic standard that can be imposed on future image production runs. It addresses need to identify correspondence between generated images and an aesthetic standard that remains constant across applications.

Methods implemented in this research follow several procedural steps. First, an abstract aesthetic dimension defined by professional artist preferences was constructed to establish an overall evaluation framework with explicit numerical values. This step would require collecting ratings for common images, then inferring statistical values for their locations on an evaluation dimension. Once established, this framework can be extended with additional images, though central core of common items defines an invariant operational construct. Successive evolutionary image streams can be compared, and it maintains aesthetic values independently of specific jury samples. By its independent nature, this aesthetic standard mimic expectation of a fully automated fitness function. Images from multiple generation runs can be compared to qualitative categories on this framework. Depending on the generating algorithm, it may require a small subset of items be embedded in every image group evaluated.

## 4.0 Method

### 4.1 Sample

This report presents results from published research [54]. The sample is 462 examinees from Johnson O'Connor Research Foundation (JOCRF) testing offices in Boston, New York, Chicago, and Dallas. Examinees were paying clients of JOCRF's aptitude-assessment service and consisted of 215 males and 247 females, predominantly white (95%), upper-middle-class, and college-educated or college-bound roughly between 15 and 40 years of age.

### 4.2 Generative image production

#### 4.2.1 Stochastic aesthetic components and algorithm

Image algorithm and production were described in prior research [53, 55]. In general, Birkhof's order-complexity ratio was invoked to establish a theoretical perspective on image development, which was implemented with a stochastic Mole-Shannon information function developed to manipulate order and complexity components independently. In this report, generated images were first validated by professional artists using rating scales, which established an aesthetic ordering of images.

#### 4.2.2 Professional artist validation

Validation of evaluation dimension In this research was conducted with a heterogeneous sample of sixty-six professional artists recruited from New York City, Chicago, San Francisco, and Dallas, Texas. Professional artists were selected after meeting inclusion criteria, which included documented history of juried awards, prizes, and commissions, working fulltime for more than five years, and evidence of formal visual arts training. Style and media were diverse as artists represented broad range of professional expertise including painters in various media, graphic artists, sculptors, photographers, architects, and so on. They provided preference judgments to presented image pairs, which were theoretically scored (0/1) where complexity was keyed correct, and their responses were transformed to an interval logit scale for dichotomously scored items.

#### 4.2.3 Image equating

Sample-dependent ratings described above, however, provide only limited foundations for an objective aesthetic dimension. Consequently, ratings were transformed to linear (equal interval) logits with absolute values, objective properties, and explicit estimate of reliability with a Rasch model. Then specification codes were identified for image locations (logits), and those scale values functioned as objective weights. Key step here was to transfer image weight to item specification code, which then functions as common item in future comparisons. Any future image generations, as well as those from modified algorithms can be evaluated on this fitness dimension without additional jury implementation. Parameterized specification codes link generated images to the evaluation dimension. In general, this overall obtained aesthetic dimension is useful for comparing aesthetic differences among separate image runs on a common scale. Applications are presented here of automatic item generation, as well as results from "seeding" figurative images with aesthetically scaled item codes rendered by an interactive artist.

Both generated images and seeded paintings were equated with linking codes selected from the evaluation dimension. A scaling constant from generated images was computed to align authentic paintings on the evaluation framework.

An overview of steps implemented to equate horizontal image runs:

- Initially, image difficulty order established by professional artists was transformed to logit values. This hierarchy established a validated evaluation dimension.
- Anchor images with specification codes were selected from evaluation dimension for equating and their aggregation established an equating constant.
- New images were generated.
- Scaling constant was added to new item pool for fitness calibration.
- Location of new items on evaluation dimension was computed based on specification codes defined previously by professional artist sample.

Figure 3 presents an elaboration of this procedure where image specification codes were selected to seed figurative paintings. In this schematic, paintings were first seeded with codes from evaluation dimension then rendered by interactive artist. Phenotype will change on multiple runs, and their location on evaluation dimension was inferred from original professional artist calibration.

### 4.3 Procedure

Generative algorithm automatically produced rule-based images that systematically varied along two dimensions, complexity and redundancy. Professional artists viewed image pairs, which were scored for conformity to theoretical model. Then rank order was transformed to logits.

Image seeds were selected after initial image equating. Image syntax in several historical Western styles were interactively `seeded` with proportions generated by the algorithm, then artist rendered. Preferences for the paintings were compared to a local sample then were equated to the evaluation dimension originally validated by professional artists using common item equating methods.

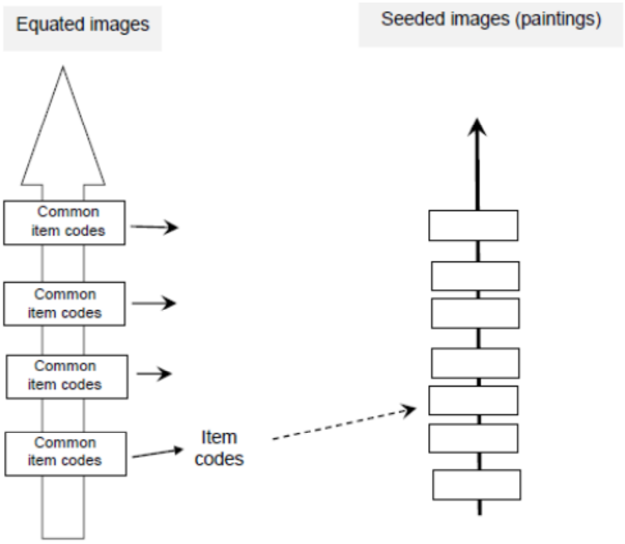


Figure 3. Interactive evolution of figurative images from “seeded” image codes. Schematic describing evolution of figurative images after “seeding”. While images shift on multiple runs, and their location is inferred from original professional artist calibration of image specification. Specifications emulated Birkoff’s order-complexity ratio variations implemented with a stochastic Mole-Shannon information function.

### 5.0 Results

Statistical analysis found complexity and redundancy in generated images accounting for 80 percent of preference variance. Figure 4 shows local preference ratings after equating to a professional artist standard, which provides uniform fitness values across multiple samples.

Figure 5 presents figurative paintings after “seeding” with images codes. While phenotypic images with identical specification will unexpectedly differ on multiple runs, their inferred stochastic locations on the evaluation dimension are fixed by empirical estimation based on original professional artist judgments.

## 6.0 Discussion

Professional artist validation model implemented here established scaffolding for a hierarchy of multi-components that accommodates image information layers. This structure can be elaborated with additional information from successive image samples. Large image samples can be equated using small sample subsets, which provide empirical values that are easily equated to the image superstructure.

Generative algorithm was validated by professional artists. Common item equating implementing Rasch logits was adapted to link local images to a standardized professional artist scale Strategy implemented in this research applied principles of aesthetic theory from empirical studies to develop a generative algorithm, and validation of an aesthetic evaluation dimension, then migration of image codes to figurative paintings. Foundations presented by complexity and order provided important insights into constructing an aesthetically valid and objective fitness evaluation.

An expedient alternative to methods presented here could implement deep learning statistical methods with a large image data base to extract features associated with aesthetic preference. Unfortunately, this approach leads to consensus scoring and ultimately ignores individual differences. Moreover, what kind of art is it? Deep learning imposes a normative standard based on computational averages to establish an arbitrary standard of aesthetic judgment. It essentially continues a separation between what laypersons and artists prefer. Not really any different from traditional dichotomy between artists and laypersons except artists are eliminated from art making and evaluation. This philosophical approach is taking visual art in a direction that is oblivious of individual differences and assumes aesthetic standards without artists. General issue of fitness evaluation for generative algorithms faces several conflicting issues.

- Contemporary cultural valuation emphasizes aesthetic novelty.
- Image properties related to semantics, expression, affective qualities, and thematic content currently not well modeled during fitness evaluation.
- Empirical aesthetic theory relies primarily on formal properties.
- Automated approaches are insensitive to skillful art production.
- Individual differences are not addressed during fitness evaluation.

.

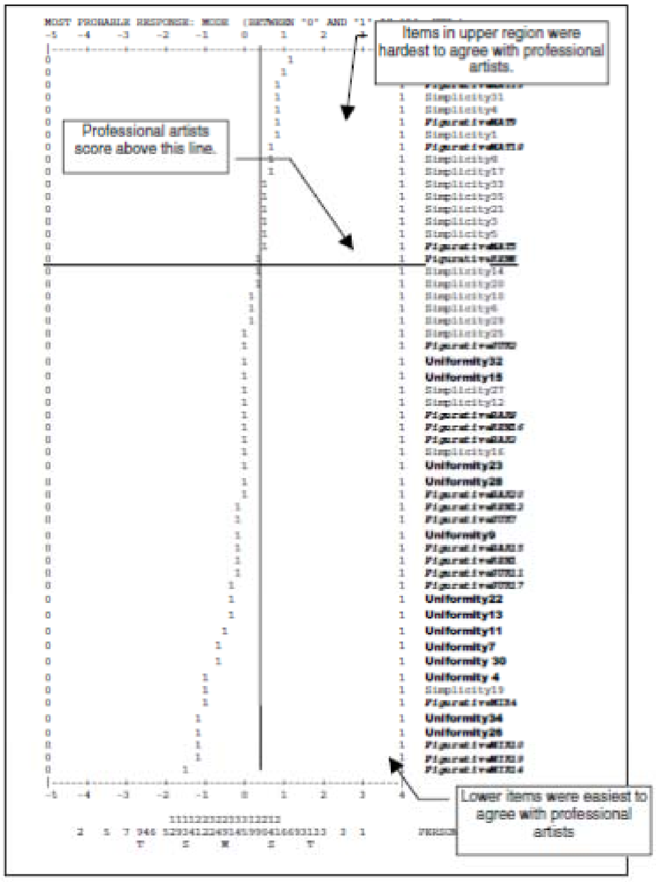


Figure 4. Local samples equated to an absolute aesthetic standard. (N=462) .

Despite obvious conceptual relevance, historical approaches in the empirical aesthetic tradition have not contributed substantially to improving fitness functions in computational aesthetics.

The most direct way to do this is via aesthetic *measure*. That is, the fitness function directly enacts some algorithmic method of scoring or ranking the aesthetic value of a specific work. This fits particularly well with aesthetic theories based around form—we will see that this is the dominant theory there too; much of the experimental work in this area explores correlations between formal aspects of visual images and the viewer's aesthetic or affective responses [56].

### 6.1 Practical implications

Fitness function based on diverse professional artists can be approximated for evolutionary art generated from small local samples. Using common item equating methods, diverse samples can be compared on a common aesthetic value framework without organizing separate evaluations.

Currently, fitness requires artist intervention. However, evaluation based on professional artists' preference that is abstracted in a mathematical scale provides an objective standard. Multiple image runs can be compared to it, and optimal images automatically selected. In addition, the initial algorithm can be changed to include properties, and those mutant images can also be compared to the professional artist standard automatically. In other words, images are evaluated individually hence not dependent on specific algorithms.

### 6.2 Future research

This report is only intended to introduce image equating and evaluation implementing psychometric methods to generative arts. Additional studies are needed to clarify specific adaptations that might make common item equating and psychometrics, in general, useful to fitness evaluation.

### 7.0 Conclusions

These results show plausibility of applying psychometric scaling methods to generative art and specifically computational aesthetic approaches that produce autonomous image runs and need an efficient and valid method of trimming lower aesthetic tails of these image distributions. Multiple thresholds can be identified on the standard, and their statistical parameters compared across samples. Moreover, this fitness framework could be implemented during image generation to guide item evolution.

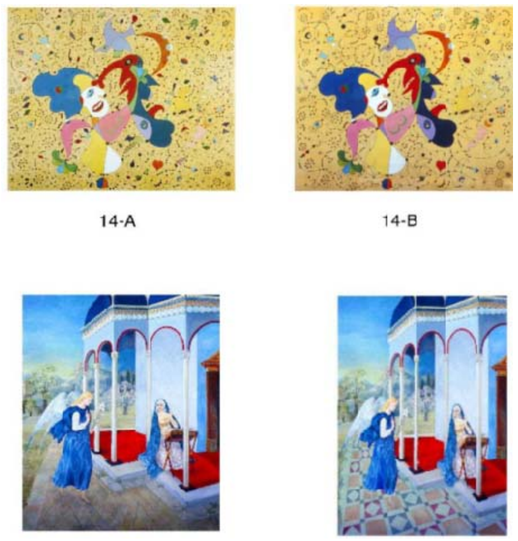


Figure 5. Seeded figurative paintings

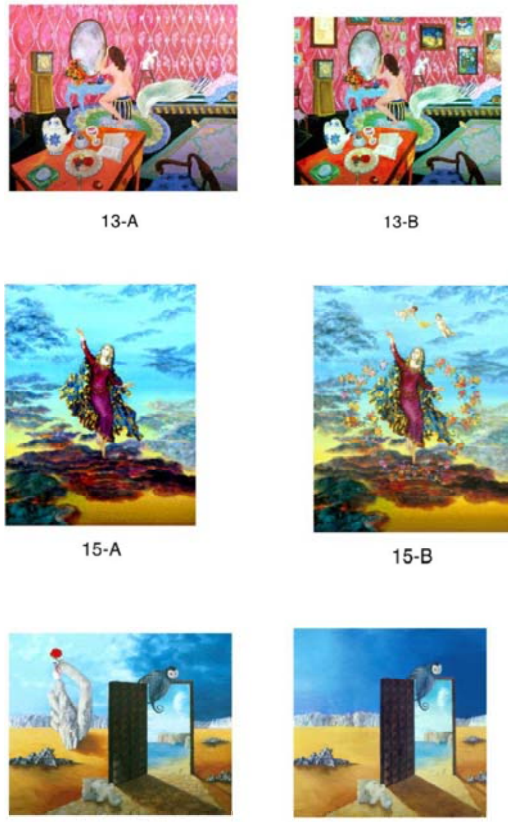


Figure 5. Seeded figurative painting (continued)

## 8.0 References

[1] Noll, A. M. (1966). Human or machine: A subjective comparison of Piet Mondrian's "Composition with Lines" (1917) and a computer-generated picture. *The Psychological Record*, *16*(1), 1-10.

[2] Zhang, K., Harrell, S., & Ji, X. (2012). Computational aesthetics: on the complexity of computer-generated paintings. *Leonardo*, *45*(3), 243-248.

[3] Zhang, K., & Yu, J. (2016). Generation of Kandinsky art. *Leonardo*, *49*(1), 48-54.

[4] Machado, P., & Cardoso, A. (2002). All the truth about NEvAr. *Applied Intelligence*, *16*(2), 101-118.

[5] Todd, Peter M., and Gregory M. Werner. 1998. "Frankensteinian Methods for Evolutionary Music Composition." In *Musical Networks: Parallel Distributed Perception and Performance*, edited by Niall Griffith & Peter M. Todd. Cambridge, MA: The MIT Press.

[6] Pei, Y. (2017). Principal component selection using interactive evolutionary computation. *The Journal of Supercomputing*, *73*(7), 3002-3020.

[7] Machado, P., Romero, J., & Manaris, B. (2008). Experiments in computational aesthetics. In *The art of artificial evolution* (pp. 381-415). Berlin, Heidelberg: Springer.

[8] den Heijer, E., & Eiben, A. E. (2010, April). Comparing aesthetic measures for evolutionary art. In *European Conference on the Applications of Evolutionary Computation* (pp. 311-320). Berlin, Heidelberg: Springer.

[9] Mureika, J. R. (2005). Fractal dimensions in perceptual color space: a comparison study using Jackson Pollock's art. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *15*(4), 043702.

[10] Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., & Sawey, M. (2011). Predicting beauty: fractal dimension and visual complexity in art. *British Journal of Psychology*, *102*(1), 49-70.

[11] Braun, J., Amirshahi, S. A., Denzler, J., & Redies, C. (2013). Statistical image properties of print advertisements, visual

artworks and images of architecture. *Frontiers in Psychology*, *4*, 808.

[12] Redies, C., Brachmann, A., & Wagemans, J. (2017). High entropy of edge orientations characterizes visual artworks from diverse cultural backgrounds. *Vision research*, *133*, 130-144.

[13] Brachmann, A., & Redies, C. (2017). Computational and experimental approaches to visual aesthetics. *Frontiers in computational neuroscience*, *11*, 102.

[14] Johnson, C. G., McCormack, J., Santos, I., & Romero, J. (2019). Understanding Aesthetics and Fitness Measures in Evolutionary Art Systems. *Complexity*, 1-14.

[15] Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2014, November). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international Conference on Multimedia* (pp. 457-466). ACM.

[16] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414-2423).

[17] Wang, Z., Chang, S., Dolcos, F., Beck, D., Liu, D., & Huang, T. S. (2016). Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155*.

[18] McCormack, J., Gifford, T., & Hutchings, P. (2019, April). Autonomy, Authenticity, Authorship and Intention in computer generated art. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)* (pp. 35-50). Cham: Springer.

[19] Lewis, M. (2008). Evolutionary visual art and design. In *The Art of Artificial Evolution* (pp. 3-37). Berlin, Heidelberg: Springer.

[20] Güçlütürk, Y., Jacobs, R. H., & Lier, R. V. (2016). Liking versus complexity: decomposing the inverted U-curve. *Frontiers in Human Neuroscience*, *10*, 112.

[21] Birkhoff, G. D. (1933). *Aesthetic Measure*. Cambridge, MA: Harvard University Press.

[22] Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., & Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, *160*, 43-57.

[23] Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *American Journal of Psychology, 103,* 53–80. http://dx.doi.org/10.2307/ 1423259

[24] Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology*, *64*, 77-107.

[25] Van Geert, E., & Wagemans, J. (2019). Order, complexity, and aesthetic appreciation. *Psychology of Aesthetics, Creativity, and the Arts, 14,* 1-21.

[26] Romero, J., Machado, P., Carballal, A., & Osorio, O. (2011, April). Aesthetic classification and sorting based on image compression. In *European Conference on the Applications of Evolutionary Computation* (pp. 394-403). Springer, Berlin, Heidelberg.

[27] Friedenberg, J., & Liby, B. (2016). Perceived beauty of random texture patterns: A preference for complexity. *Acta Psychologica*, *168*, 41-49.

[28] Gauvrit, N., Soler-Toscano, F., & Guida, A. (2017). A preference for some types of complexity comment on "perceived beauty of random texture patterns: A preference for complexity". *Acta Psychologica*, *174*, 48-53.

[29] Bauerly, M., & Liu, Y. (2006). Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. *International Journal of Human-Computer Studies*, *64*(8), 670-682.

[30] al-Rifaie, M. M., Ursyn, A., Zimmer, R., & Javid, M. A. J. (2017, April). On symmetry, aesthetics and quantifying symmetrical complexity. In *International Conference on Evolutionary and*

*Biologically Inspired Music and Art* (pp. 17-32). Cham: Springer.

[31] Attneave, F. (1959). Stochastic composition processes. *The Journal of Aesthetics and Art Criticism*, *17*(4), 503-510.

[32] Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing*, *3*(2), 236-252.

[33] Romero, J., Machado, P., Santos, A., & Cardoso, A. (2003, April). On the development of critics in evolutionary computation artists. In *Workshops on Applications of Evolutionary Computation* (pp. 559-569). Berlin, Heidelberg: Springer.

[34] Fechner, G. T. (1876), Vorschule der Ästhetik. (Introduction to aesthetics). Leipzig: Breitkopf & Härtel.

[35] Thurstone, L. L. (1927) The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384-400.

[36] Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529-54.

[37] Thurstone, L. L. (1954). Measurement of Values. *Psychological Review*, *61*, 47.

[38] Eysenck, H. J. (1940). The general factor in aesthetic judgements. *British Journal of Psychology*, *31*(1), 94.

[39] Eysenck, H. J. (1941). The empirical determination of an aesthetic formula. *Psychological Review*, *48*(1), 83.

[40] Eysenck, H. J. (1941). Type-factors in aesthetic judgements. *British Journal of Psychology*, *31*(3), 262.

[41] Johnson O'Connor Research Foundation. (1990). *Artistic Judgment Project II: Construct Validation*. Technical Report 1990-1). Chicago: Author. (ERIC No. 017 064).

[42] Corradi, G., Belman, M., Currò, T., Chuquichambi, E. G., Rey, C., & Nadal, M. (2019). Aesthetic sensitivity to curvature in real objects and abstract designs. *Acta Psychologica*, *197*, 124-130.

[43] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

[44] Platt, J. R. (1961). Beauty: Pattern and change (402-430). In D. W. Fiske and S. R. Maddi (Eds.), *Functions of Varied Experience*. Homewood IL: Dorsey Press.

[45] Moles, A. (1968). Information Theory and Esthetic Perception. *Journal of Aesthetics and Art Criticism,* 26, 552-554.

[46] Berlyne, D. E. (1974). *Studies in the New Experimental Aesthetics: Steps Toward an Objective Psychology of Aesthetic Appreciation*. New York: Hemisphere.

[47] Berlyne, D. E. (1971). *Aesthetics and Psychobiology.* New York: Appleton-Century-Crofts.

[48] Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

[49] Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359-374.

[50] Bezruczko, N., & Schroeder, D. H. (1987). *Visual Designs Test: Abstract Items.* Chicago: Johnson O'Connor Research Foundation, Inc.

[51] Johnson O'Connor Research Foundation. (1991). *Artistic Judgment III: Artist Validation. Technical Report 1991-1.* Chicago: Author.

[52] Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling and Linking. Methods and Practices*. Third edition. New York: Springer-Verlag.

[53] Bezruczko, N. (2014). Automatic item generation implemented for measuring artistic judgment aptitude. *Journal of Applied Measurement.* 15, 1-25.

[54] Bezruczko, N. (2002). A multi-factor Rasch scale for artistic judgment. *Journal of Applied Measurement*. *3*(4), 360-399

[55] Bezruczko, N. (2013). Generative Art Simplifies Psychometrics of Artistic Judgment Aptitude. *Generative Art*, 168. Accessed on October 18, 2019 at https://generativeart.com/

[56] Johnson, C. G., McCormack, J., Santos, I., & Romero, J. (2019). Understanding Aesthetics and Fitness Measures in Evolutionary Art Systems. *Complexity*, *2019*.