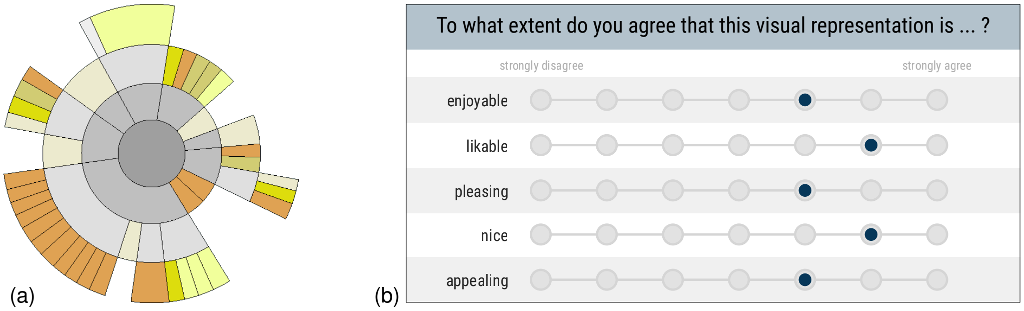
BeauVis: A Validated Scale for Measuring the Aesthetic Pleasure of Visual Representations

Tingying He, Petra Isenberg, Raimund Dachselt, and Tobias Isenberg



To what extent do you agree that this visual representation is ... ?

	strongly disagree						strongly agree
enjoyable	○	○	○	○	●	○	○
likable	○	○	○	○	○	●	○
pleasing	○	○	○	○	●	○	○
nice	○	○	○	○	○	●	○
appealing	○	○	○	○	●	○	○

Fig. 1: For (a), one participant's data (b) on our BeauVis scale in its recommended version; (a) from [20], © IEEE, used with permission.

Abstract—We developed and validated a rating scale to assess the *aesthetic pleasure* (or *beauty*) of a visual data representation: the BeauVis scale. With our work we offer researchers and practitioners a simple instrument to compare the visual appearance of different visualizations, unrelated to data or context of use. Our rating scale can, for example, be used to accompany results from controlled experiments or be used as informative data points during in-depth qualitative studies. Given the lack of an aesthetic pleasure scale dedicated to visualizations, researchers have mostly chosen their own terms to study or compare the aesthetic pleasure of visualizations. Yet, many terms are possible and currently no clear guidance on their effectiveness regarding the judgment of aesthetic pleasure exists. To solve this problem, we engaged in a multi-step research process to develop the first validated rating scale specifically for judging the aesthetic pleasure of a visualization (osf.io/fxs76). Our final BeauVis scale consists of five items, "enjoyable," "likable," "pleasing," "nice," and "appealing." Beyond this scale itself, we contribute (a) a systematic review of the terms used in past research to capture aesthetics, (b) an investigation with visualization experts who suggested terms to use for judging the aesthetic pleasure of a visualization, and (c) a confirmatory survey in which we used our terms to study the aesthetic pleasure of a set of 3 visualizations.

Index Terms—Aesthetics, aesthetic pleasure, validated scale, scale development, visual representations.

1 Introduction

Visualization as a field relies on many foundations, including computer science, mathematics, human-computer interaction, psychology, social sciences, design, and art. The study of aesthetics is essential to several of these foundations and, subsequently, visualization. Yet, aesthetics is an elusive concept or phenomenon that is subjective and potentially socially constructed [61]. It is a vast research field with whole research institutes dedicated to its subfield empirical aesthetics,[1] which studies "how people experience, evaluate, and create objects aesthetically" [16]. In visualization research, aesthetics has mostly been studied in terms of a visualization's visual appeal or beauty. This focus is often described under the term *aesthetic pleasure* or *aesthetic experience* in the psychology literature. In this paper, we focus on the concept of *aesthetic pleasure*, rather than the entire concept of aesthetics.

Aesthetic pleasure is an important aspect of visualizations. It has been suggested to affect the usability and effectiveness of a visualization [20, 37] and has the potential to communicate [15] and engage viewers [2, 68]. To make empirically-grounded statements about the impact of aesthetic pleasure on visualization use, however, we first need a set of research instruments to study this concept. Fechner [16] posited that aesthetic pleasure can be studied just like other forms of perception and proposed to analyze study participants' reactions to certain stimuli. Such methods require participants to order or *rank* objects based on aesthetic preference or to *rate* them according to a degree of preference [53]. Based on these original ideas, researchers have developed rating scales to study the aesthetic pleasure of websites [44, 52] or objects [12]. Rating scales are measurement instruments that consist of a group of rating items later combined into a composite score. These rating scales are typically used to indicate levels of an underlying phenomenon (called latent variable or construct) that are hard to observe by direct means [26]. For the study of aesthetic pleasure, these rating scales complement the toolbox of methods such as brain scans, eye tracking, or in-depth qualitative methods by being easy to deploy and analyze.

Yet, while scales have been developed in other domains, we lack validation to know whether these approaches also work to study the aesthetic pleasure of visualizations in particular or if other or new terms are required. Instead, researchers currently pick their own terms to evaluate aesthetic pleasure and ask participants to rate visualizations according to, for example, how "visually appealing" [1], "elegant" [27], or "aesthetic" [39] they are. Unfortunately, without a validated instrument we cannot be certain that these ad-hoc approaches to understanding aesthetic pleasure are reliable and sufficient. In addition, the abundance of terms used in the literature makes it difficult to compare results. To address this limitation, we developed and validated a scale specifically for measuring the aesthetic pleasure of visual data representations, i. e., the images resulting from a visualization process [72, 73].

With our work we provide a simple validated instrument for researchers and practitioners to assess and compare the aesthetic pleasure

- *Tingying He (何汀滢), Petra Isenberg, and Tobias Isenberg are with Université Paris-Saclay, CNRS, Inria, LISN, France. E-mail: {tingying.he | petra.isenberg | tobias.isenberg}@inria.fr.*
- *Raimund Dachselt is with Technische Universität Dresden, Germany. E-mail: raimund.dachselt@tu-dresden.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

[1] Such as the Max Planck Institute of Empirical Aesthetics in Germany or the Penn Center for Neuroaesthetics in the USA.

of different representations. Our scale cannot be used for measuring people's impressions of the visual representation that relate to data—such as memorability, intuitiveness, informativeness, or understandability or context-of-use related aspects such as appropriateness. We validated our scale to capture first impressions, without interactivity and context. We do not mean to replace in-depth qualitative analyses of aesthetic experience or other methods of empirical analysis. Our validated scale can be used, however, *together* with other approaches, to deliver another data point or to help create hypotheses that may explain other empirical results. Beyond the final scale itself, our work makes several contributions: First, we conducted a systematic review of how aesthetic pleasure has been studied in the literature and extracted a set of terms used in the visualization literature. Next, we conducted surveys with 31 visualization experts, who we asked for additional terms. We narrowed down our combined set of 209 terms to 37 terms and asked experts to rate them according to their relevance to the construct of aesthetic pleasure. We then derived a final set of 3–5 terms from a crowd-sourced experiment in which we had 1001 participants rate 15 different visualizations using a subset of the expert-rated terms. Finally, we conducted another confirmatory crowd-sourced analysis of 3 visualizations in which participants used our 5-item scale to rate the visual data representations' aesthetic pleasure.

2 Related Work

As we already noted, aesthetics is an elusive concept that does not have a universally accepted definition. Generally speaking, aesthetics is related to beauty and its appreciation. In this section, we start by defining aesthetic pleasure and then summarize empirical aesthetic methods. Next, we present past work on the study of aesthetics in the field of visualization and finally, we review how researchers in related fields measured aesthetic pleasure.

2.1 Definition of Aesthetic Pleasure

The debate about whether beauty is subjective or objective has persisted throughout history. Reber et al. [61] summarized that, in the philosophical tradition, there are three main ways of looking at beauty. According to the *objectivist view*, beauty is a characteristic of an object that causes a delightful experience in any appropriate perceiver. Several features of an object can contribute to its aesthetics, such as balance, symmetry, clarity, etc. According to the *subjectivist view*, in contrast, anything can be beautiful. Beauty depends on perceivers, and all attempts to discover the rules of beauty are futile. The most modern approach is an *interactionist view* that combines the previous two views and regards beauty as the function of both the characteristics of the object and the perceiver. We adopt this interactionist view in our work.

In the past, researchers have used "beauty" and "aesthetic pleasure" interchangeably. For instance, Reber et al. [61] defined beauty as "a pleasurable subjective experience that is directed toward an object and not mediated by intervening reasoning" and equate it to the concept of aesthetic pleasure, meaning essentially the same thing. This definition also fits well with how many researchers (e. g., [20, 21, 35, 70]) approached the concept in visualization, and we adopt this definition to describe the construct we want to measure in our scale. We can see similar definitions in other work, e. g., "the pleasure people derive from processing the object for its own sake, as a source of immediate experiential pleasure in itself, and not essentially for its utility in producing something else that is either useful or pleasurable" [28], but see this definition as largely equivalent to the first one, which we adopt.

Aesthetic pleasure is part of the concept of aesthetic experience as it is used in empirical aesthetics, which can be understood as the experience that arises from a unique combination of cognitive and emotional processes [45]. Aesthetic appreciation consists of three main modes [60]: aesthetic pleasure, emotions evoked by an artwork, and understanding of an artwork. Our work focuses on the aesthetic pleasure of visualizations, so it is to study the first modes of aesthetic appreciation. Graf and Landwehr [32, 33] proposed a comprehensive model of aesthetic pleasure called the Pleasure-Interest Model of Aesthetic Liking. This model shows that there are two forms of processing aesthetics, resulting in different forms of liking: *automatic processing* and *controlled processing*. Automatic processing is driven by a stimulus, which is a quick and instinctive judgment based on pleasure or displeasure as a response to the stimulus, and leads to *pleasure-based liking*. Controlled processing is driven by the perceivers, which leads to *interest-based liking*. This model involves both the stimuli and perceiver, so it is in line with our interactionist view on beauty.

2.2 Empirical Aesthetics

There are two main ways to study aesthetics [54]. *Philosophical aesthetics*, with a long tradition starting in ancient Greece, uses a top-down approach, examining general concepts and then applying them to specific cases. *Empirical aesthetics*, established by Gustav Theodor Fechner in the 19th century, works bottom-up, examining specific cases (e. g., what people like or dislike about something) and then deriving a set of principles from them. In our work, we mostly follow the approach of empirical aesthetics as we use empirical methodologies [54].

Experimental aesthetics is one of the most essential subfields of empirical aesthetics. It generally relies on the measurement of historical data, verbal ratings and judgments, measurement of nonverbal behavior, and measurement of psychophysiological changes. Among these methods, the one most relevant to our own work is the measurement of verbal responses. Researchers use this method to collect some aspect of the way participants experience a stimulus. Most commonly, participants are asked to provide "descriptive aspects of the stimuli (e. g., their complexity, regularity, or novelty), evaluative aspects of the hedonic value (e. g., degree of interest or pleasure, liking, beauty, or attractiveness), and internal states (e. g., evoked emotions or meanings)" [54]. Verbal ratings can, thus, be recorded and analyzed in several ways, but a common approach is to establish a scale that targets the construct described by the participants—which is what we do in this work.

2.3 Aesthetic Pleasure in Visualization

The term *aesthetics* is often used in visualization to describe a property of a visual representation that is separated from how understandable, informative, or memorable it is; and that instead focuses on its beauty or visual appeal. In this way the concept aligns with the definition we adopted for *aesthetic pleasure*, and we set out to study it in more detail.

In 2005, Chen [21] listed the study of "pretty or visually appealing" visualization designs under the heading of aesthetics as one of the top ten unsolved problems in information visualization. Since then, however, research dedicated to visualization aesthetics has been sparse, perhaps due to the challenges of describing, measuring, and quantifying aesthetics [70]. Lau and Vande Moere [43] proposed *information aesthetics* as a term that describes aesthetics in the context of visualization as a construct meant to augment "information value and task functionality." Vande Moere and Purchase [70], later, equate aesthetics with attractiveness in their work on the role of design in information visualization but describe aesthetics as a concept that is broad and includes aspects such as "originality, innovation, and novelty" [70]. The authors specifically call for research that aims to explain the reasons for aesthetic experiences. This is specifically NOT something our rating scale will accomplish. Our scale will allow researchers to compare the aesthetic pleasure of visual data representations as it is judged by participants, but it will not allow us to explain *why* participants rated the representation in a certain way. To derive reasons for aesthetically pleasurable experiences or to establish a comprehensive aesthetic measurement the scale can, however, be included in larger questionnaires or in qualitative studies (interviews, observations, etc.).

Aesthetics has also been regarded as an important factor in some subfields of visualization. For example, aesthetics has been identified as a heuristic for evaluating ambient visualization [48]. Also, within graph drawing, specific aesthetics heuristics have been defined as properties of a graph that not only describe attractiveness but impact readability and understanding [7, 59]. These include aesthetics related to symmetry, edge lengths, or the minimization of edge crossings. These heuristics have also been extended, e. g., to aesthetics heuristics for dynamic graph visualization [5] or the faithfulness criterion [55] based on readability.

Several studies have been conducted by previous researchers for *evaluating* the aesthetics of a visualization. Much of this work has borrowed from methods introduced many years ago in empirical aesthetics; e. g., the use of rating scales. Cawthon and Vande Moere [19] presented a conceptual model for assessing aesthetics as part of an information visualization's user experience. In another study [20], they asked participants to rate visualizations on a scale from "ugly" to "beautiful" to judge their aesthetics. Many other scales have been used in visualization. For example, Harrison et al. [35] used a rating scale from "not at all appealing" to "very appealing" in their study on infographics. Ajani et al. [1] used a rating scale from "very hideous" to "very beautiful" in their study on the aesthetics of three visualization designs. Chen et al. [22] used a rating scale from "nice" to "ugly" to study the aesthetic appearance of visualization technique. These examples target what we call aesthetic pleasure but are mostly based on intuition rather than a verified instrument that can ascertain that the terms indeed measure the aesthetic pleasure of visualizations reliably and validly. Also, compared with a multi-item scale, one item lacks enough information to calculate psychometric properties such as reliability [31] and leads to less accurate results due to item-specific measurement error [13, 31].

2.4 Measuring Aesthetic Pleasure outside of Visualization

In the field of HCI, researchers have developed several validated scales to measure the aesthetic appreciation of websites and interactive products. These scales were developed and validated broadly following a standard process which we outline in Sect. 3.

To measure the aesthetic pleasure of websites, Lavie and Tractinsky [44] proposed a scale with two dimensions, which they termed *classical aesthetics* and *expressive aesthetics*. The *classical aesthetics* dimension comprises the five items "clean," "clear," "pleasant," "symmetrical," and "aesthetic." The *expressive aesthetics* dimension, in contrast, includes the five items "original," "sophisticated," "fascinating," "creative," and "uses special effects." Moshagen and Thielsch [52], however, pointed out that Lavie and Tractinsky's scale has the following problems: the items "symmetrical" and "uses special effects" are not necessarily aesthetic judgments, it is hard to explain why the term "aesthetic" only relates to the classic aesthetic dimension, and their items are too abstract to be used for improving the design. Based on Lavie and Tractinsky's scale, Moshagen and Thielsch thus proposed a scale with the four dimensions of simplicity, diversity, colorfulness, and craftsmanship, with items such as "the layout appears well structured," "the design appears uninspired," "the color composition is attractive," and "the layout appears professionally designed."

To measure aesthetic pleasure for designed artifacts, Blijlevens et al. [12] pointed out that previous scales do not measure aesthetic pleasure separately from its determinants. Hence, they proposed the Aesthetic Pleasure in Design Scale in which they distinguish between both. Their scale includes five items: "beautiful," "attractive," "pleasing to see," "nice to see," and "like to look at." In addition, they also pointed out some dimensions suitable for measuring prominent determinants of aesthetic pleasure such as typicality, novelty, unity, and variety.

In addition to scales specific to aesthetics, some scales for user experience also include dimensions related to aesthetics. The widely used AttrakDiff questionnaire [36], e. g., includes *hedonic quality* and *overall attractiveness*, which are related to aesthetic pleasure and include items such as "pleasant," "attractive," and "creative." The User Experience Questionnaire (UEQ) [65] has a dimension *attractiveness* to capture the overall impression of a product, with items such as "enjoyable," "good," and "friendly." The meCUE questionnaire [51] has a dimension *visual aesthetics*, with items such as "creatively designed," "attractive," and "stylish." These questionnaires, however, should be administered after full exposure to a product to measure people's experience—different from our goal of capturing viewers' first impressions.

To the best of our knowledge, there exists no targeted scale yet for measuring the aesthetic pleasure of visual data representations. Until now, visualization researchers can only use scales that are designed for interactive products in general; for example, the AttrakDiff questionnaire has been used in several visualization studies (e. g., [17, 76]).

3 The BeauVis Scale: Methodology Overview

We largely followed the process described by DeVellis and Thorpe [26] and Boateng et al. [13] to establish a validated scale of *aesthetic pleasure* for future use in the visualization field. This process contains four steps: (1) generating a pool of possible terms, (2) item review, (3) item evaluation, and (4) scale validation.

At the start of our work, we decided to target a Likert scale [46] response format, with equally weighted items. We also pre-determined to use a 7-point Likert scale throughout our work with the same categories for each item, from 1 = strongly disagree to 7 = strongly agree—except for Survey 2 in which we ask about the relevance of terms, for which a lower number is encouraged [26]. We chose an odd number of response categories to offer participants a neutral rating and the number 7 to strike a balance between discriminability and usability; in addition, the related literature on aesthetic pleasure scales also uses 7-point Likert scales facilitating comparison. However, our final scale could certainly be used with a larger or smaller number of response categories.

We began our research by investigating past visualization publications for their use of terms relating to some form of aesthetic ratings, such as in evaluations of techniques or tools. We also checked the literature for terms used in aesthetics-related scale development in other related fields as additional input. As a final source of candidate terms we conducted a survey among visualization experts for terms they would suggest to use. We then narrowed down the aggregated list of terms based on several objective criteria, and again asked visualization experts to rate how important each of the remaining terms was for studying aesthetic pleasure in visualization. This gave us a list of 31 terms, which we then used in a crowd-sourced experiment that asked participants to rate 15 diverse visual data representations with respect to each of the final terms. We then conducted an exploratory factor analysis and calculated the reliability of scales with a smaller number of items. Based on these analyses, we arrived at our final five-item BeauVis scale. Finally, we conducted another crowd-sourced experiment to validate our final scale using a confirmatory factor analysis, calculated Cronbach's alpha, convergent validity and discriminant validity. We will discuss our detailed approach next.

4 Generating a Pool of Possible Terms

The first step in our process was the generation of a pool of terms that could describe the construct of *aesthetic pleasure*. We drew these possible items from the literature and experts.

4.1 Literature Review

Our literature review involved two sources: the VIS literature as a source of terms used in the past by the community as well as related work on scales in other domains as a source of terms considered and used for measuring the same construct (i. e., aesthetic pleasure).

Collecting terms from the visualization literature: To determine which terms the community had used in the past to study aesthetics, we reviewed IEEE VIS papers (1991–2020) and TVCG and CG&A journal papers presented at IEEE VIS (2011–2021)—3 189 paper PDF files in total. We extracted the text of these files and searched for the occurrence of "aesthetic," "likert," "questionnaire," and "interview." We retrieved 1 061 articles with at least one of our four search terms, and then summarized the results in a spreadsheet (recording publication year, journal, paper title, DOI link, found search term, and PDF filename). The first author then opened each of these PDFs and checked whether the authors had indeed conducted a study that recorded participants' subjective feelings about the aesthetics of a visual data representation. We focused on collecting terms used as part of rating scales. We found terms in 68 papers, but many did not relate to aesthetic pleasure. For example, we did not include terms that were used to judge interaction, usability, or task-related aspects (e. g., how confident a participant felt in their answers). We included, however, terms that described an aesthetic-related subjective feeling such as "clarity" or "understandability." With this initially rather broad spectrum of terms, we accounted for the complexity of the aesthetic construct and ensured that we would not miss any potentially relevant terms.

Term grouping, adjective forming, and counting. To be able to better analyze the use of terms by the visualization community, we wanted to count terms which in turn required extensive cleaning and rechecking of the literature. We turned all terms into adjectives and merged different forms of the same word. For example, we merged "understandable," "understandability," and "ease of understanding" all into "understandable." In addition, we went back to the 68 papers to verify the counts and checked the context of each term to determine what these terms measured (e. g., visual encoding, design, interface, etc.). Based on the latter analysis, we kept all terms that measured a visual encoding (e. g., visualization technique, representation, design etc.) but discussed among the authors cases that measured interface, tool, or layout. We could not completely disregard this last group because many of the tools described in the visualization literature are visual analysis tools, which, in turn, naturally comprise visual representations as a major component; so an aesthetic-related assessment of such a tool may also largely be an evaluation of the visual representation(s) included within. We thus based our decision on our impression if the evaluation related to the visual representation (included), as opposed to the interaction or usability (excluded). After completing this step, we retained a final list of 41 adjective terms. The most common terms were aesthetic (20×), understandable (12×), and intuitive (9×).

Term categorization. Next, we tagged the 41 terms with the types of judgments they target: aesthetic, emotion-oriented, cognitive-oriented, data-aesthetic, or other. Terms could receive more than one tag. We considered a term to make an *aesthetic judgment* if it clearly applied to the aesthetic pleasure caused by a visual representation. The most common terms in this category were "aesthetic" (20×), "well-designed" (5×), and "cluttered" (5×, cross-tagged with cognitive-oriented). *Emotion-oriented judgments* describe broad emotional or affective reactions to visuals. The most common terms in this category were "pleasing" (7×), "engaging," "enjoyable," and "likable" (all 4×). We categorized terms as targeting *cognitive-oriented judgments* when they seemed to primarily assess the cognitive process of understanding or analyzing data with the visualization. The most common terms in this category were "understandable" (12×), "intuitive" (9×), and "clear" (7×). Fourth, terms targeting *data-aesthetic judgments* are those whose aesthetic judgment hinged largely on the combination of data and design. We tagged only three terms in this category "expressive" (4×, cross-tagged with aesthetic), "informative" (4×, cross-tagged with cognitive-oriented), and "suitable" (1×). Four terms seemed to target *another judgment*, such as being related to quality ("high-quality," 1×), innovation ("innovative," 2×), or established practice ("conventional," 2×). The most common word with more than one tag was "cluttered" (5×), which can be considered to make both an aesthetic and a cognitive-oriented judgment. We show the final list and classification in Table 6 in the appendix.

Term input from related fields: In addition to reviewing visualization literature, we also consulted literature from related fields about aesthetic pleasure scales. We found four scales for assessing the aesthetics of websites and interactive products that are most aligned with our own goals or had high citation counts [12, 36, 44, 52]. We extracted the terms studied in these four papers to compare them to the ones we had collected. For two of these papers [12,44] we were able to extract all terms that the authors had considered in the development of their scale from the papers. For a third [52], the authors kindly e-mailed us their early list of considered terms (not included in their final paper) and we translated these German terms into English. From the fourth paper [36] we could only use the terms the authors selected as their final scale. For all terms from these four papers we followed the same cleaning and tagging process as before for the visualization literature and then combined them with our list. The total list from our literature review thus included 176 terms (Table 7 in the appendix).

4.2 Expert Suggestion—Survey 1

To supplement our literature review, next we conducted a pre-registered (`osf.io/wvehs`) and IRB-approved (Inria COERLE, avis № 2022-12) survey to ask for expert input on words we had not yet considered.

Participants. We invited 57 visualization experts among a wide spread of topic expertise to participate in our survey by direct e-mail. We selected participants based on our knowledge of their work and their reputation in the visualization community. Participants were not compensated for taking part in the study. After sending the invitation e-mails, we waited for one week and, during this time, received 31 complete responses (9 female, 21 male, 1 gender not disclosed; past experience in visualization research: mean = 19.7 years). All responses were valid and we included them in our analysis.

Procedure. We first asked participants to complete the informed consent form and to answer background questions about their gender and expertise. We then explained the study scenario and task which involved wanting to investigate people's subjective opinions about the aesthetics of a visualization they had created, using a 7-point Likert scale with the question: "To what extent do you agree or disagree with the following statement: This visualization is [. . .]." We then asked each of our expert participants to provide us with at least three words they would want to use or could envision to use for filling the blank in the question. We gave them the opportunity to leave additional comments after providing us with their term suggestions.

Results. From the 31 completed surveys we collected 113 different words. We cleaned these words by removing duplicates, fixing typos, as well as merging them and forming adjectives as before. Through this process we received 77 unique adjectives (Table 8 in the appendix) and counted their frequencies. The most common terms were: "beautiful" (18×), "pleasing" (16×), and "aesthetic" (15×). We then combined these terms with the terms we collected from the literature and categorized them as before. Through this process our list of terms added 33 new terms and grew to a total of 209 terms (Table 9 in the appendix).

5 Term Filtering

As a next step we needed to select a meaningful subset of the 206 terms we had identified, so that we would have a manageable number to administer to a development sample (Sect. 6). We thus first removed less relevant terms based on several considerations (Sect. 5.1), followed by an expert review via a second survey (Sect. 5.2).

5.1 Filtering on Occurrence and Semantics

After several rounds of discussions among the author team and consulting the literature on scale development [13, 26], we settled on the following criteria to decide whether we should retain a term or not.

1. The terms needed to be **related to *aesthetic pleasure*** rather than *understanding* or *comprehension* of a visual representation or its data (e. g., we excluded "informative," "clear," or "confusing").
2. The terms had to have **appeared at least twice** in one of the three resources we used for our item generation: visualization papers, other relevant aesthetics scale papers, or expert suggestions.
3. The terms should be **usable in a rating scale** and have a **clearly good or bad connotation** (e. g., we excluded "complex" because a complex aesthetic could be seen as positive or as negative).
4. The terms should be **easy to understand** (e. g., we excluded "consistent" because it would be unclear according to what aspect a visual appearance would be consistent) and their **interpretation should be clear** (e. g., we excluded "novel" because it would require people to know what "old" visualizations look like; we also excluded "drab" as a rare term that is not easily understood by many non-native speakers of English).
5. The terms had to **clearly apply to an assessment of a visual representation** (e. g., we excluded "dynamic" because, within visualization, the term may be read as referring to the property of being animated or interactive, rather than a dynamic aesthetic).
6. The terms should **not be pairs of opposite adjectives**. We only retained negative terms that did not have a clear positive opposite (e. g., we excluded "ugly" as the opposite of "beautiful").

Based on the first criterion, we excluded terms that made a cognitive judgment because, for such a judgment, one needs to understand the data and we aimed to assess the visuals only. We had an intensive deliberation about terms that made an emotional judgment. We finally decided to include them because such a judgment can be closely related to the aesthetic *pleasure* generated by a visual representation and it can be difficult to separate those terms from emotion-only expressions. In the Pleasure-Interest Model of Aesthetic Liking [32, 33], the interest could be considered as an aesthetic emotion [60]. Thus, the boundary between aesthetic pleasure and aesthetic emotion is not always clear. Ultimately, we thus arrived at a shortlist of 37 terms (see Table 10 in the appendix) that we categorized as making an aesthetic, emotional, and other judgment, that served as the input for an expert review.

5.2 Expert Review—Survey 2

Next, we conducted a second pre-registered (`osf.io/5gmut`) and IRB-approved (Inria COERLE, avis № 2022-12) survey to elicit expert feedback on the relevance of the 37 terms for measuring the aesthetic pleasure of a visual data representation.

Participants. We e-mailed the same experts (excluding one who had participated in a pilot, for a total of 56 experts), and received 25 complete responses after three days (8 female, 16 male, 1 gender not disclosed; past experience in visualization research: mean = 20.1 years). All responses were valid and we included them in our analysis.

Procedure. We first asked the participants to provide their informed consent and background information. We then introduced them to our definition of aesthetic pleasure and asked them to rate "how relevant do you think the following terms are for judging or describing the aesthetic pleasure of a visualization?". The rating scale included 5 points from 1 being 'not at all relevant' to 5 being 'very relevant.' Finally, we again allowed them to leave additional comments.

Results. For each term, we calculated the median and mode of all participants' answers. From the 37 total terms, 32 terms received a mode of 3 or above or a median of 3 or above. Among these 32 terms, we removed the term "aesthetic" based on our own discussion and the recommendation of one expert, as we feared the term to be too abstract and elusive to rate reliably. We thus arrived at a final list of 31 terms (Table 11 in the appendix) that we used in our exploratory phase.

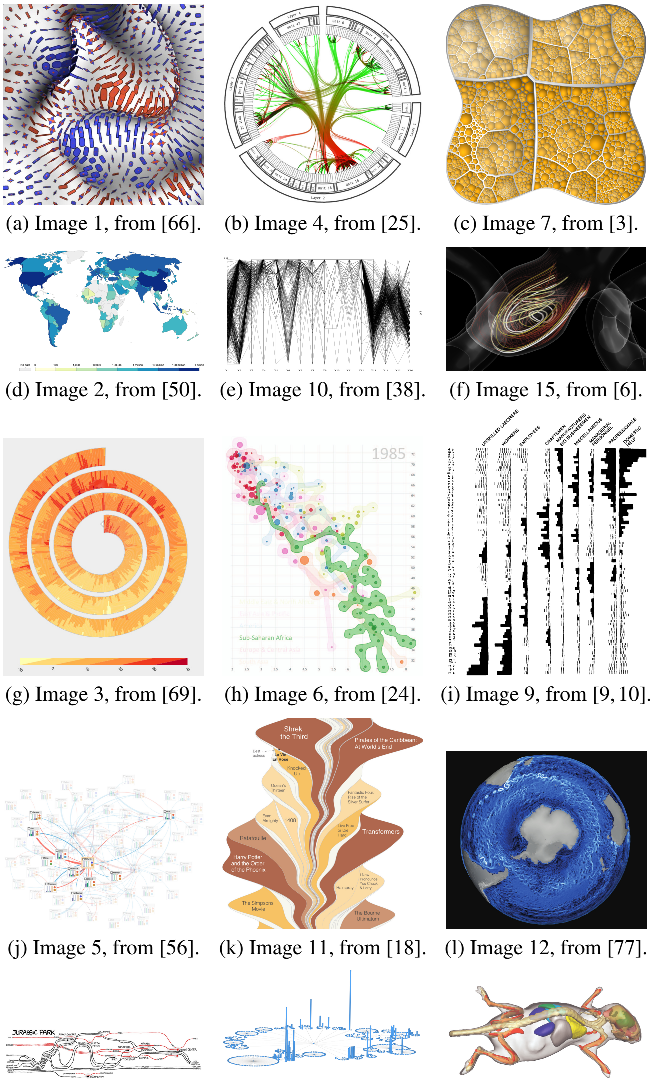
6 Exploratory Phase: Exploratory Factor Analysis

During scale development, it is important to establish how a set of items actually studies the targeted construct, aesthetic pleasure in our case. Specifically, it is important to establish whether the ratings for the terms we collected are all caused by the same property of aesthetic pleasure or perhaps multiple identifiable factors of aesthetic pleasure such as symmetry, clarity, or familiarity. So we needed to identify the minimum number of these hypothetical factors as a next step of our analysis [75]. In addition, 31 terms are too many for the easy-to-administer research instrument we were targeting. We thus needed to identify the terms that performed best and exclude terms that did not perform well. Exploratory factor analysis (EFA) [75] has specifically been developed as an analytic tool to help researchers with these challenges. To generate data for an EFA we conducted a third pre-registered (`osf.io/az8sm`) and IRB-approved (Inria COERLE, avis № 2022-12) survey, in which participants used our 31 terms to rate a set of visualizations.

6.1 Exploratory Survey—Survey 3

Stimuli. In total, we selected 15 representative images that showed a variety of different visualization techniques that participants would rate. For our selection of specific visual representations (Fig. 2) we used different criteria that may affect aesthetic pleasure judgments. We wanted to cover a wide variety of areas of visualization work and different approaches to visualizations designs, such as 2D/3D, black vs. white backgrounds, abstract vs. physical content, hand-crafted vs. computer-generated aesthetic, and black and white vs. colorful. All images came from scientific publications, because our scale targets research evaluations such as surveys.

Participants. There is no consensus about sample size for factor analysis but general recommendations say that the more items to test, the more participants are required. In line with two suggestions [8, 13] we targeted a sample size of 200 participants per visualization. We recruited participants through Prolific, who had to be fluent English speakers and to be of legal age (18 years in most countries). Participants received a compensation of € 10.2 per hour.



(a) Image 1, from [66]. (b) Image 4, from [25]. (c) Image 7, from [3].

(d) Image 2, from [50]. (e) Image 10, from [38]. (f) Image 15, from [6].

(g) Image 3, from [69]. (h) Image 6, from [24]. (i) Image 9, from [9, 10].

(j) Image 5, from [56]. (k) Image 11, from [18]. (l) Image 12, from [77].

(m) Image 8, from [47]. (n) Image 13, from [49]. (o) Image 14, from [41].

Fig. 2: The 15 visual representations that we used as examples from the visualization literature in our analysis. Image permissions: (a–c, e, h, k–l, o) © IEEE; (d) © Springer-Nature; (f) © Wiley; (g) © C. Tominski and H. Schumann; (i) © EHESS [10, p. 230, #3]; (j) © ACM/Nobre et al. [56]; (n) by Marai et al. [49], 🅭🅯 CC-BY 4.0; (m) by R. Munroe (originally XKCD # 657), 🅭🅯🄏 CC-BY-NC 2.5. All images are used with permission from the respective copyright holders.

Procedure. We first asked the participants to provide their consent and collected demographics. Then we asked them to rate 3 visualizations, randomly selected from the 15 visualizations. They rated each visualization according to the question "To what extent do you agree or disagree with the following statement: The visualization is ...." For each of the 31 terms, we asked participants to choose an answer on a 7-point Likert item ranging from "strongly disagree" to "strongly agree." We showed the terms and visualizations in a random order, because we could not counter-balance the order due to the limitations of the Limesurvey system we used. We showed the images without captions so that participants would focus on the visuals. We also included one attention check question for each visualization. We asked participants to answer the online survey on a computer or laptop due to the high number of items to rate and the visual length of the scale.

6.2 Results

We recruited a total number of 1001 participants, who all provided their informed consent. We excluded 2 participants who each answered our survey twice due to a technical error. We also excluded 10 par-

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209390

ticipants who answered two or three of our attention check questions incorrectly. We used the remaining 989 responses for our analysis (ages: mean = 28.3, SD = 9.4; 389 female, 589 male, 11 gender not disclosed; education: 618 Bachelor or equivalent, 138 Master's or equivalent, 22 PhD or equivalent, 211 other) and reversed their scores for the negative term "cluttered." Due to our random assignment of participants to images, each image was rated by approx. 200 people (mean = 197.7, SD = 19.5, min = 178, max = 218).

6.3 Exploratory Factor Analysis (EFA)

We followed Watkins' systematic guide to EFA [75] and implemented all tests using the psych R package [62], applying them separately for each visual representation.

Appropriateness of EFA. Before conducting the EFA, we needed to confirm whether our data was suitable for EFA. First, we calculated a correlation matrix of all terms for each of the 15 visualizations. Only "provoking" and "cluttered" had a low correlation (< 0.3) with other terms, for all 15 visualizations. The other correlations were outside the interval $[-0.3, .3]$, which meant that the data was suitable for EFA. We then conducted Bartlett's test of sphericity [4]. The results showed that $p < .001$ for all 15 visualizations, which indicates that there is a large-enough correlation between terms. We also conducted a Kaiser-Meyer-Olkin (KMO) test [40]. All individual terms' KMO values were above 0.7. Based on all these tests, we confirmed that our data's correlation matrices were factorable and then submitted them to EFA.

Extracting Factors. We conducted an exploratory factor analysis of the 989 responses to the 31 terms for each image. We chose a common factor analysis model rather than PCA (principal component analysis) as it is recommended for the creation of measurement instruments such as rating scales [29, 75]. Roughly speaking, common factor analysis targets to find hypothetical factors that *caused* the ratings of participants, while PCA components are *defined* by the ratings.

We used *scree plots* and *parallel analysis* (for details on both see DeVellis and Thorpe's book [26]) to determine the potential factors of our scale. Parallel analysis, which uses purely statistical criteria to determine the number of factors, indicated that there was more than one factor for all 15 visualizations (Table 1). We complemented this objective finding with a more subjective analysis using scree plots. Here, we inspected the scree plots for all images such as the one shown in Fig. 3. We noted that, in all plots, the eigenvalues of the second factor were close to 1, similar to the pattern seen in Fig. 3 (we show all plots in Appendix B). The eigenvalues represent how much information is captured by a factor. If a factor's eigenvalue is 1, it captures the same proportion of information as a single item [26]. As we were after the compression of our item pool, we decided that factors that captured only little more information than single items would not be retained. We thus conducted our EFA for all images using one factor only. However, to not overlook a potentially prominent factor, we also conducted an exploratory analysis using an EFA for two factors using a Varimax (orthogonal) and Promax (oblique) rotation and analyzed the data (we provide the data of this analysis in Appendix G). For a few images, we analyzed how the terms were split into two factors but were unable to extract meaningful factor descriptions. Therefore, we confirmed that our items indeed measured one factor (aesthetic pleasure) and based our further analysis on the results of the EFA with one factor only.

Reducing Terms. The next step in scale development is to find an acceptable number of final terms to use. One of the important outputs of an EFA is a table with factor loadings per term. The higher a factor loading, the more the term defines the factor or, in our case, the better it is able to describe aesthetic pleasure. Based on their factor loadings, the terms the least descriptive for aesthetic pleasure in our data were "provoking" and "cluttered" with factor loadings below 0.5 for all of the 15 visualizations, see Fig. 4. Twelve terms had a factor loading of > 0.7 for all of 15 visualizations, which are considered high values [34]. In decreasing order of their average factor loadings these were: "likable, pleasing, enjoyable, appealing, nice, attractive, delightful, satisfying, pretty, beautiful, lovely, and inviting." We removed all other terms and did not further consider them in the creation of our final scale.

At this point we had 12 terms left, which we could combine into even smaller scales. For each possible scale one can compute a reliability statistic that indicates whether a scale would perform in consistent and predictable ways. A perfectly reliable scale would always consistently measure the true aesthetic pleasure of a visual representation. Reliability measures approximate this "true" value by computing the proportion of a "true" score to the observed score. We used Cronbach's alpha as our reliability measure, which looks at the scale's total variance attributable to a common source and which is the most commonly used measure of reliability in scale development [26].

Table 1: Number of factors as output by the parallel analysis.

Image	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Factors	2	2	3	3	2	2	3	2	3	2	2	3	2	2	2

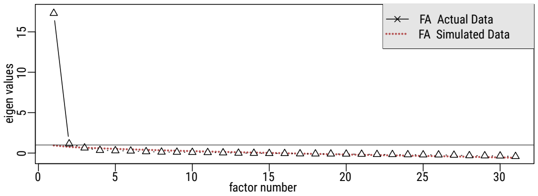
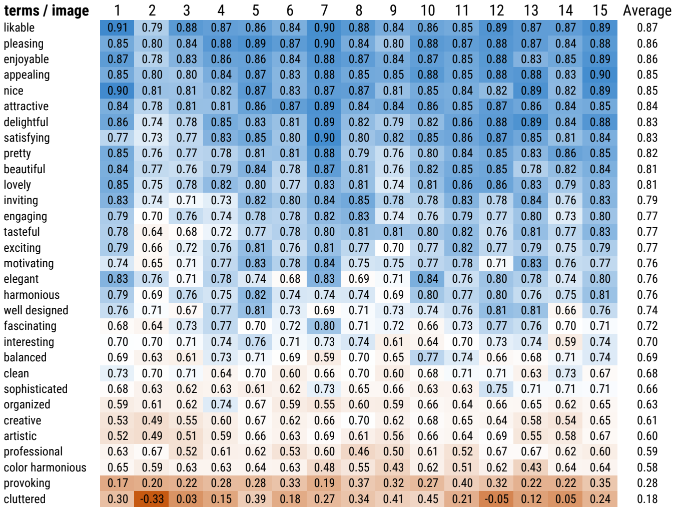


Fig. 3: Scree plot for Image 1 (3D surface glyphs).



terms / image	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average
likable	0.91	0.79	0.88	0.87	0.86	0.84	0.90	0.88	0.84	0.86	0.85	0.89	0.87	0.87	0.89	0.87
pleasing	0.85	0.80	0.84	0.88	0.89	0.87	0.90	0.84	0.80	0.88	0.87	0.88	0.87	0.84	0.88	0.86
enjoyable	0.87	0.78	0.83	0.86	0.86	0.84	0.88	0.87	0.84	0.87	0.85	0.88	0.83	0.85	0.89	0.86
appealing	0.85	0.80	0.80	0.84	0.87	0.83	0.88	0.85	0.85	0.88	0.85	0.88	0.88	0.83	0.90	0.85
nice	0.90	0.81	0.81	0.82	0.87	0.83	0.87	0.87	0.81	0.85	0.84	0.82	0.89	0.82	0.89	0.85
attractive	0.84	0.78	0.81	0.81	0.86	0.87	0.89	0.84	0.84	0.86	0.85	0.87	0.86	0.84	0.85	0.84
delightful	0.86	0.74	0.78	0.85	0.83	0.81	0.89	0.82	0.79	0.82	0.86	0.88	0.89	0.84	0.88	0.83
satisfying	0.77	0.73	0.77	0.83	0.85	0.80	0.90	0.80	0.82	0.85	0.86	0.87	0.85	0.81	0.84	0.83
pretty	0.85	0.76	0.77	0.78	0.81	0.81	0.88	0.79	0.76	0.80	0.84	0.85	0.83	0.86	0.85	0.82
beautiful	0.84	0.77	0.76	0.79	0.84	0.78	0.87	0.81	0.76	0.82	0.85	0.85	0.78	0.82	0.84	0.81
lovely	0.85	0.75	0.78	0.82	0.80	0.77	0.83	0.81	0.74	0.81	0.86	0.86	0.83	0.79	0.83	0.81
inviting	0.83	0.74	0.71	0.73	0.82	0.80	0.84	0.85	0.78	0.78	0.83	0.78	0.84	0.76	0.83	0.79
engaging	0.79	0.70	0.76	0.74	0.78	0.78	0.82	0.83	0.74	0.76	0.79	0.77	0.80	0.73	0.80	0.77
tasteful	0.78	0.64	0.68	0.72	0.77	0.78	0.80	0.81	0.81	0.80	0.82	0.76	0.81	0.77	0.83	0.77
exciting	0.79	0.66	0.72	0.76	0.81	0.76	0.81	0.77	0.70	0.77	0.82	0.77	0.79	0.75	0.79	0.77
motivating	0.74	0.65	0.71	0.77	0.83	0.78	0.84	0.75	0.75	0.77	0.78	0.71	0.83	0.76	0.77	0.76
elegant	0.83	0.76	0.71	0.78	0.74	0.68	0.83	0.69	0.71	0.84	0.76	0.80	0.78	0.74	0.80	0.76
harmonious	0.79	0.69	0.76	0.75	0.82	0.74	0.74	0.74	0.69	0.80	0.77	0.80	0.76	0.75	0.81	0.76
well designed	0.76	0.71	0.67	0.77	0.81	0.73	0.69	0.71	0.73	0.74	0.76	0.81	0.81	0.66	0.76	0.74
fascinating	0.68	0.64	0.73	0.77	0.70	0.72	0.80	0.71	0.72	0.66	0.73	0.77	0.76	0.70	0.71	0.72
interesting	0.70	0.70	0.71	0.74	0.76	0.71	0.73	0.74	0.61	0.64	0.70	0.73	0.74	0.59	0.74	0.70
balanced	0.69	0.63	0.61	0.73	0.71	0.69	0.59	0.70	0.65	0.77	0.74	0.66	0.68	0.71	0.74	0.69
clean	0.73	0.70	0.71	0.64	0.70	0.60	0.66	0.70	0.60	0.68	0.71	0.71	0.63	0.73	0.67	0.68
sophisticated	0.68	0.63	0.62	0.63	0.61	0.62	0.73	0.65	0.66	0.63	0.63	0.75	0.71	0.71	0.71	0.66
organized	0.59	0.61	0.62	0.74	0.67	0.59	0.55	0.60	0.59	0.66	0.64	0.66	0.65	0.62	0.65	0.63
creative	0.53	0.49	0.55	0.60	0.67	0.62	0.66	0.70	0.62	0.68	0.65	0.64	0.58	0.54	0.65	0.61
artistic	0.52	0.49	0.51	0.59	0.66	0.63	0.69	0.61	0.56	0.66	0.64	0.69	0.55	0.58	0.67	0.60
professional	0.63	0.67	0.52	0.61	0.62	0.53	0.60	0.46	0.50	0.61	0.52	0.67	0.67	0.62	0.60	0.59
color harmonious	0.65	0.59	0.63	0.63	0.64	0.63	0.48	0.55	0.43	0.62	0.51	0.62	0.43	0.64	0.64	0.58
provoking	0.17	0.20	0.22	0.28	0.28	0.33	0.19	0.37	0.32	0.27	0.40	0.32	0.22	0.22	0.35	0.28
cluttered	0.30	-0.33	0.03	0.15	0.39	0.18	0.27	0.34	0.41	0.45	0.21	-0.05	0.12	0.05	0.24	0.18

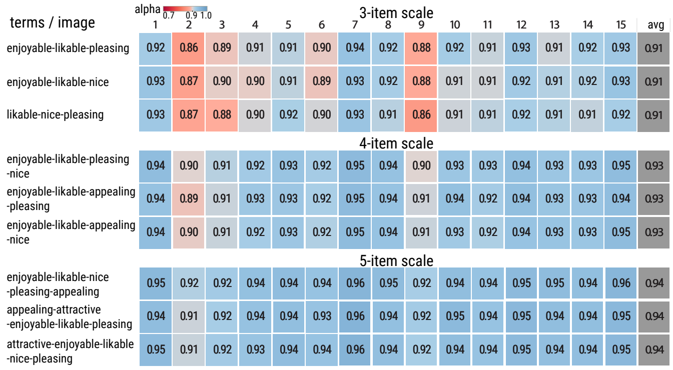
Fig. 4: Factor loadings for all 31 terms and images using diverging red–blue color scale centered at 0.7, which is mapped to white.

Because we were aiming for a lightweight instrument, we tested the reliability of final scales of size 3–5. Three items is the minimum number for the statistical identification of a factor and four to six items per factor have been recommended [30]. Here, choosing the right size is a tradeoff between usability and reliability. Cronbach's alpha increases with the number of items, but more items require participants to spend more time to answer and rate visual representations. We calculated Cronbach's alpha for all potential 3-item, 4-item and 5-item combinations of these 12 high factor loading terms, for all 15 visual representations that we started to use in Sect. 6.1 (i. e., those in Fig. 2).

Final Scale. The reliability of scales constructed through the combinations of the highest factor-loading terms was high overall (Fig. 5) and multiple word combinations are possible.

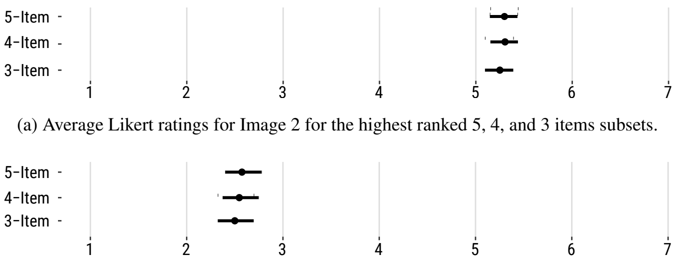
The best 3-item subset (enjoyable, likable, pleasing) had an alpha of 0.91 (range of 0.86–0.93 for the images tested), the 4-item subset (enjoyable, likable, pleasing, nice) had a reliability of 0.93 (range of 0.9–0.95), and the 5-item subset (enjoyable, likable, pleasing, nice, appealing) a reliability of 0.94 (range of 0.92–0.96). In Fig. 5 we see that alpha generally rises with more items. To further understand the effect of a 3-, 4-, or 5-item subset we conducted an exploratory analysis in which we calculated the average aesthetic ratings for each image as if participants had only used those items. These calculations are exploratory because we cannot guarantee that the presence of additional items did not influence the ratings of our participants (yet to exclude

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209390


alpha
0.7
0.9
1.0

terms / image	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
3-item scale																
enjoyable-likable-pleasing	0.92	0.86	0.89	0.91	0.91	0.90	0.94	0.92	0.88	0.92	0.91	0.93	0.91	0.92	0.93	0.91
enjoyable-likable-nice	0.93	0.87	0.90	0.90	0.91	0.89	0.93	0.92	0.88	0.91	0.91	0.92	0.91	0.92	0.93	0.91
likable-nice-pleasing	0.93	0.87	0.88	0.90	0.92	0.90	0.93	0.91	0.86	0.91	0.91	0.92	0.91	0.91	0.92	0.91
4-item scale																
enjoyable-likable-pleasing -nice	0.94	0.90	0.91	0.92	0.93	0.92	0.95	0.94	0.90	0.93	0.93	0.94	0.93	0.93	0.95	0.93
enjoyable-likable-appealing -pleasing	0.94	0.89	0.91	0.93	0.93	0.92	0.95	0.94	0.91	0.94	0.92	0.94	0.93	0.93	0.94	0.93
enjoyable-likable-appealing -nice	0.94	0.90	0.91	0.92	0.93	0.92	0.95	0.94	0.91	0.93	0.92	0.94	0.93	0.93	0.95	0.93
5-item scale																
enjoyable-likable-nice -pleasing-appealing	0.95	0.92	0.92	0.94	0.94	0.94	0.96	0.95	0.92	0.94	0.94	0.95	0.95	0.94	0.96	0.94
appealing-attractive -enjoyable-likable-pleasing	0.94	0.91	0.92	0.94	0.94	0.93	0.96	0.94	0.92	0.95	0.94	0.95	0.94	0.94	0.95	0.94
attractive-enjoyable-likable -nice-pleasing	0.95	0.91	0.92	0.93	0.94	0.94	0.96	0.94	0.92	0.94	0.94	0.95	0.94	0.94	0.95	0.94

Fig. 5: Cronbach's alpha for each image on the most reliable 3-, 4-, and 5-item subsets of the remaining 12 terms with factor loading > 0.7.



(a) Average Likert ratings for Image 2 for the highest ranked 5, 4, and 3 items subsets.

(b) Average Likert ratings for Image 9 for the highest ranked 5, 4, and 3 items subsets.

Fig. 6: Comparison of ratings from subsets of the rating items for Image 2 and Image 9 that had the lowest and highest average ratings in our image set. We show the plots for the other images in the appendix.

these possible effects we conducted a confirmatory factor analysis in the next step described in Sect. 7). Fig. 6 shows, for two images, that there were only small variations in the average ratings. The average rating of all 15 images (see Fig. 43–57 in the appendix) also reflects the balance of the aesthetic quality of the images we selected: the number of images scoring above and below the middle score were almost equal.

We thus conclude that a combination of 3, 4, and 5 items would produce reliable results. Scales with Cronbach's alpha > 0.7 are considered reliable [13], so even our minimum 3-item scale was reliable. Nonetheless, we recommend using the 5-item scale for its even higher reliability and because it can still be completed quickly by participants.

7 Validation Phase

The final scale development step is to validate the developed scale. Broadly speaking, a validated scale should actually measure the construct (aesthetic pleasure) and should do so reliably. We conducted a confirmatory factor analysis (CFA) to test the scale's dimensionality—i. e., we checked whether we indeed measure just one factor of aesthetic pleasure as planned during the exploratory phase (Sect. 6) [13]. Then we tested the reliability of the results on new data we collected. Finally, we determined several measures of the construct validity of our scale that target how well the scale measured aesthetic pleasure.

7.1 Validation Survey—Survey 4

For this phase we conducted a fourth pre-registered (osf.io/gsq6p) and IRB-approved (Inria COERLE, avis № 2022-12) survey, like the last one also using crowd-sourcing. Again, participants rated visualization but this time using the 5-item scale proposed in the previous section. To validate our results we had participants rate 3 visualizations that had been previously assessed for aesthetic pleasure by other researchers (and participants) using a different measuring instrument [20].

Stimuli. We chose to partially reproduce findings from Cawthon and Vande Moere's experiment on the effect of aesthetics on visualization usability [20]. They had asked participants to assess the aesthetic pleasure of 11 visualizations using a one-item 100-point scale from "ugly" to "beautiful." To achieve a broader range of aesthetic experience, we selected three (SunBurst, StarTree, and BeamTree, see Fig. 7) out of the 11 visualization techniques that were rated to be the most "beautiful" (Sunburst), most "ugly" (BeamTree), and somewhat neutral (StarTree).

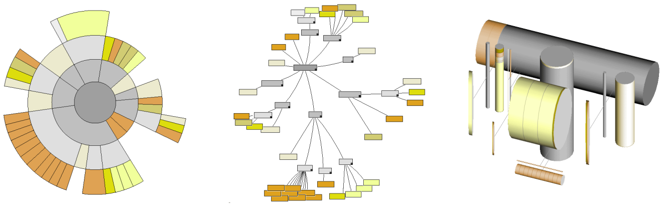


Fig. 7: The visual representations SunBurst, StarTree, and BeamTree from Cawthon and Vande Moere's [20] study of perceived aesthetics that we used in our validation. SunBurst (left) was ranked as most beautiful, StarTree (middle) as neutral, and BeamTree (right) as most ugly in the experiment [20]. All images are © IEEE, used with permission.

Cawthon and Vande Moere kindly provided their stimuli images to us, and we used them as stimuli in our validation survey. We hypothesized that our BeauVis scale would rank these visualizations similarly from high to low as follows: SunBurst, StarTree, and BeamTree.

Participants. We targeted to recruit 200 participants from the general public on Prolific, using the same approach as in Survey 3 (Sect. 6.1).

Procedure. We also followed the same procedure as we did in Survey 3, which we described in Sect. 6.1, with the following exceptions: We used a clear within-subjects design where all participants rated all three visual representations (SunBurst, StarTree, BeamTree) with the five terms in our scale (enjoyable, likable, pleasing, nice, appealing) as well as with Lavie and Tractinsky's [44] 5-item scale for measuring classic aesthetics of websites (aesthetic, pleasant, clear, clean, symmetric) (see Sect. 2.4). We used this additional five-item scale for validating convergent validity, which we explain below. We only used one attention check question in this survey.

7.2 Results

We recruited a total number of 201 participants. All participates provided their informed consent. We excluded 4 participants who answered the attention check questions incorrectly. We used the remaining 197 responses for our analysis (ages: mean = 25.1, SD = 6.4; 69 female, 126 male, 1 gender not disclosed; education: 125 Bachelor or equivalent, 22 Master's or equivalent, 2 PhD or equivalent, 48 other). Participants received a compensation of € 10.2 per hour.

Confirmatory Factor Analysis (CFA) is a statistical technique that allows us to make inferences about the constructs that were measured. As aesthetic pleasure was the single construct we targeted during the exploratory phase, we used CFA to examine the construct structure as well as to verify the number of constructs measured and the item-construct relationships via factor loadings, similar to the earlier EFA. We used the methods based on structural equation modeling (SEM), which is the most commonly used CFA method [26]. We evaluated model fit by means of a series of statistical tests in CFA, including χ^2, Tucker Lewis Index (TLI), Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR), and Root Mean Square Error of Approximation (RMSEA). We implemented all tests using the lavaan R package [63], applying them separately for each image.

Goodness of Fit. To calculate how well the scale items describe the aesthetic pleasure construct we needed to define a model that describes our only factor (aesthetic pleasure) defined as the sum of the five items of our scale. In Table 2 we can see that almost all indices show a good fit of this model to the data. For the three visual representations, virtually all of the following criteria are met that are indicative of a good fit [13]: χ^2 is not significant, TLI ≥ 0.95, CFI ≥ 0.95, SRMR ≤ 0.08. The only value that does not meet these criteria is the p-value of the χ^2 test for BeamTree, but this statistical test can be sensitive to the size of the sample and should not be used as the basis for accepting or rejecting a scale [64, 71]. For a robust assessment using this test one would have needed participant pools of N ≥ 400 [14] or even N $\geq 2\,000$ [79]. The RMSEA values of SunBurst and StarTree are ≤ 0.06—also indicative

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209390

Table 2: Goodness of fit indices (TLI = Tucker Lewis Index; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation).

	SunBurst	StarTree	BeamTree
p-value (χ^2)	0.290	0.222	0.016
TLI	0.998	0.996	0.982
CFI	0.999	0.998	0.991
SRMR	0.009	0.011	0.014
RMSEA	0.034	0.045	0.095

Table 3: Standardized factor loading for five items, for each image.

Item	Factor Loading		
	SunBurst	StarTree	BeamTree
enjoyable	0.893	0.878	0.911
likable	0.914	0.925	0.874
pleasing	0.889	0.895	0.893
nice	0.845	0.877	0.888
appealing	0.910	0.842	0.889

Table 4: Cronbach's alpha for each visualization.

	SunBurst	StarTree	BeamTree
Cronbach's Alpha	0.95	0.946	0.95

Table 5: Pearson correlation.

	SunBurst	StarTree	BeamTree
Classic Aesthetic	0.84	0.88	0.87
Age	0.07	0.12	0.14

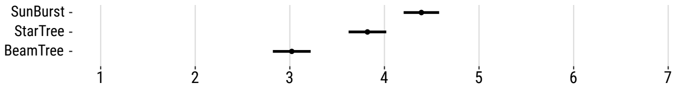


Fig. 8: Average results with our scale of the three visualization.

of a good fit [13]. The RMSEA value of BeamTree is 0.095, which is considered to be sufficient as RMSEA values $\in [.05, .10]$ suggest "acceptable" fits [42]. Based on the above results, we can say the CFA results validated our one-factor model of the BeauVis scale.

Factor Loadings. Factor loadings describe the correlation between the items and the aesthetic pleasure factor. Values close to 1 indicate that the construct of aesthetic pleasure strongly influences the item ratings. In the SEM approach of CFA, standardized factor loading values of ≥ 0.7 indicate a well-defined model [34]. As we show in Table 3, the values for all 5 items in our scale are well above 0.7.

In summary, the CFA confirmed the one-factor structure of our scale and that the items in the scale are well able to measure the construct.

Reliability: As before, we assessed the reliability of the scale using Cronbach's alpha for each image. As we show in Table 4, all alpha scores are well above 0.7 and thus our scale can be considered reliable.

Validity: A scale is considered to be valid if it can be established that it indeed measures the construct it was developed for [13]. The validity of a scale should not only be ensured at the end of the scale development phase but also throughout the earlier phases of the process [13]. According to scale development theory [13, 26], the validity of our scale can be determined according to three main aspects:

Content validity is the degree to which aesthetic pleasure is indeed reflected by the terms we chose for the scale. To establish content validity, the main method is to ask experts who are familiar with aesthetic pleasure of visualizations to review the initial item lists. We did so early in the process as explained in Sect. 5.2.

Criterion validity looks at whether the scale can explain or predict another criterion related to the "performance" of a visualization. For example, we could theoretically assess connections between aesthetic pleasure and a visualization's usability or memorability. Practically, however, establishing whether such a connection exists would require established and validated ways to measure usability or memorability of visualizations and much more complex research setups. We, therefore, did not test for criterion validity.

Construct validity describes how well a scale indeed is related to and measures the concept it promises to assess. To assess it we focused on three indices of construct validity: *convergent validity*, *discriminant validity*, and *differentiation by known group*.

The first, *convergent validity*, refers to whether different ways of measuring the same construct yield similar results. It can be demonstrated by a high correlation between a newly developed scale with other scales that promise to measure the same or a closely related construct [13]. To assess convergent validity we had participants rate visualizations also using Lavie and Tractinsky's [44] scale for assessing the aesthetic of websites. We chose their scale's *classic aesthetic factor* because its items ("aesthetic," "pleasant," "clear," "clean," and "symmetric") are more suitable for assessing visual representations than the items of their *expressive aesthetic factor*. The latter includes the term "uses special effects," e. g., which is hard to interpret for our static images. For a high convergent validity our scale's results should be correlated with those of Lavie and Tractinsky's classic aesthetics scale. As we show in Table 5, we found that, indeed, the Pearson correlation between both scales were high (i. e., > 0.5), for all three visualizations.

Second, *discriminant validity* allows us to understand to which degree a new scale measures a unique concept and that it is not related to other variables to which it should not be related. We can check for this validity by testing the correlations between the newly developed scale and other, existing measures.[2] In our case there is no reason to assume that the participant's age would be related to aesthetic pleasure and we thus use age for establishing discriminant validity, in line with Lavie and Tractinsky's [44] work. As shown in Table 5, the Pearson correlation factors between our scale and age for the three visual representations were low (i. e., well below 0.3), so we can conclude that our scale has at least discriminant validity concerning age.

Finally, in our last analysis of validity we look at the *differentiation by known groups*. Here, our "groups" are the three visualizations from Cawthon and Vande Moere (Fig. 7) [20] for which we have empirically established aesthetic measures. To contribute to construct validity we then compared the results of our scale to their previous scores to check if the scores were as expected and that the new scale could discriminate between the aesthetic pleasure of the three visualizations [13]. In Fig. 8 we show the average results for these three visual representations for the five items of our scale, with a 95% confidence interval. The scores, from highest to lowest, are SunBurst, StarTree, and BeamTree, which fully align with Cawthon and Vande Moere's results. In Cawthon and Vande Moere's original study the individual aesthetic ranking result for SunBurst was 58%, StarTree was 49% (estimated from Fig. 4 in [20]), and BeamTree was 36%. We translated these results into our 7-point Likert scale through a linear mapping, the result for SunBurst was 4.48 ($= 1 + (7-1) \cdot 0.58$), the result for StarTree was 3.94 ($= 1 + (7-1) \cdot 0.49$), and the result for BeamTree was 3.16 ($= 1 + (7-1) \cdot 0.36$). As one can see in Fig. 8, these results are sufficiently close to the actual scores in our survey such that we can also conclude validity w.r.t. differentiation by known groups.

8 Discussion and Limitations

In this section we discuss the use of our BeauVis scale, reflect on the terms they include, and discuss limitations and future work.

8.1 Guidelines for and Limits of Using the Scale

The BeauVis scale provides a simple instrument to *compare* the aesthetic pleasure of different visual representations. The mean of all items can be used to obtain a single value [57]. This value, however, should be seen in comparison and not be interpreted as an absolute measurement of how beautiful an image is or whether it is "sufficiently"

[2] Note that, essentially, we would need to check for this lack of correlation to an infinite amount of other measures, yet here we follow the established procedure [13] and the examples from the literature (e. g., [44]).

beautiful. Nor does the scale establish an exhaustive or final measurement of the broad concept of aesthetics. Some experts in our two expert surveys mentioned that aesthetics cannot be measured. This is a valid opinion representing subjectivist views of aesthetics that attributes the experience entirely to the viewer (Sect. 2.1). We address this view somewhat by narrowing our scale toward "aesthetic pleasure" or "beauty," rather than the full concept of aesthetics. Our scale can be used alone to quickly compare the aesthetic pleasure of two representations or together with other test results and be interpreted carefully in-context. Cawthone and Vande Moere [20], e. g., used an aesthetic pleasure rating in their larger study on aesthetic pleasure and user experience. Xu et al. [78] studied the effectiveness (in terms of time and error) of representations but also asked people for their aesthetic preferences to compare techniques. Stusak et al. [67] conducted a primarily qualitative study on data physicalizations but also asked participants to rate their aesthetics on a Likert scale to accompany the wealth of other data collected.

When the BeauVis scale should be administered in a study, however, requires careful consideration. We validated the scale by asking participants to rate visualizations without having interacted with them and without having read the data; that is, we asked for their first impressions. As such, we recommend to use our scale at the beginning of an empirical study similar to how we did in our own experiment. Administering a visualization rating scale after an experiment, however, is common practice and here results need to be interpreted in light of usage experiences or the data content. We addressed the concern of a possible difference between pre- and post-study administration somewhat by excluding terms related to comprehension of the visualized data. Yet, further formal validation should establish potential differences.

8.2 The Rating Question

In setting up the scale we had to decide on a rating question and settled on "To what extent do you agree that this visual representations is...?" We debated the wording of this question deeply and decided to use one that would not require clear opposing terms to be established, such as "ugly" vs. "beautiful," because we found it difficult to find suitable opposites for many terms (e. g., "likable," "pleasing," etc.). Our chosen rating question also required all terms to be adjectives, which is not always easy to achieve. When we first asked experts to suggest terms, some experts criticized our statement as they found the question to constrain suitable terms. Changes in the question might certainly make other terms possible but would also require some of our terms to be changed and the scale to be re-evaluated. Nevertheless, we expect small changes in the question not to have a great effect on the results. The term "visual representation," which we used to focus on the visual artifact and not the process of its creation [72, 73], could be exchanged by the name of the actual technique being studied, for example.

8.3 Terms in Our Scale

All terms of our final scale are related and similar to one-another. In a unidimensional, one-factor scale like ours all items measure the same construct. Their similarity stems from the reliability calculation that determines correlations. Having some similarity is useful: by having five terms in our final scale, we address variations of people's understanding of the individual terms and reduce noise. Other terms that we originally tested, in the end, turned out not to be descriptive of the concept of aesthetic pleasure and were removed.

Apart from "nice," all other terms came from what we had labeled the "emotion" category, despite the fact that there was a larger number of terms we tested in the "aesthetic" category. Clear outliers in our term exploration were "provoking" and "cluttered," but the terms "color-harmonious, professional, artistic, creative, and organized" also generally had low factor loadings for all images. In retrospect, this makes sense as many of these terms require viewers to assess the visual representation according to something else that may or may not be known. To assess whether a visual representation looks professional or artistic, e. g., one needs to know what an amateur version of it would look like. Such comparisons are not needed for terms like "pleasing" or "enjoyable," which can be answered through purely personal experience.

The terms in our scale relate to other scales of aesthetic pleasure, but have small differences. The Aesthetic Pleasure in Design scale [12], for instance, also contains the terms "pleasing to see, like to look at, and nice to see" in addition to "beautiful", and "attractive." And Lavie and Tactinsky's scale for websites [44] includes "pleasant design" under the factor classic aesthetics. Our items are specific to visualization in that they avoid terms that require a cognitive assessment of the visual representation and how understandable the data was. We purposefully avoided, e. g., terms such as "clear" that are included in Lavie and Tactinsky's scale. In addition, we avoided terms that may be important for aesthetic product ratings but less important for the aesthetic pleasure of visual representations. "Innovative," e. g., may be important for products and is a term in the AttrakDiff scale [36], but it is difficult to judge in a visualization context where participants would need to know a "standard" visual representation to rate the innovation of a new one.

We debated for a long time but finally eliminated terms that were not clearly positive or negative when applied to visual representations such as "simple" or "complex." These terms can certainly describe what a visual representation looks like but would not be able to clearly measure aesthetic pleasure because there are certainly both beautiful and ugly "simple" data representations. By avoiding terms that can describe aesthetic pleasure in two different ways the combined result of all items in the scale is more comparable.

9 Conclusion and Future Work

With our BeauVis scale we provide visualization researchers with an instrument to compare the aesthetic pleasure of the visual representations they create. With its combination of five descriptive terms it allows collect reliable average results when compared to using just a single elusive term such as "aesthetics" or even "aesthetic pleasure" itself [13, 31]. As we followed a standard scientific procedure for scale development, our approach can also serve as an example for the visualization community to establish further validated scales.

Our scale can certainly be used to compare the aesthetic pleasure within a single experiment. To compare between experiments it would require the scale to be administered using the same questions, ratings, and items but also comparable contextual factors. Preceding questions, prior use, different user groups, or motivations can all influence the scale responses [26]. As such, future work on understanding the differences of scale responses based on contextual factors would be very valuable. Other future work includes establishing related scales for certain subfields of visualization. For example, graph drawing already has a set of dedicated aesthetic criteria which should be considered in term collection for this research area. Other scales could target related constructs such as the aesthetics of the interaction with a visualization tool or the emotional experience with an artifact (e. g., [11, 74]). Naturally, our own scale can and should also further be validated, such as by conducting a Test-Retest to assess participants' consistency across time.

Acknowledgments

We thank all visualization experts who responded to our Surveys 1 and 2, whose input was an essential foundation of our work. We also thank Morten Moshagen and Meinald T. Thielsch for their early terms list [52] and Nick Cawthon and Andrew Vande Moere for their high-quality study images [20]. This work was partly supported by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC-2068 – 390729961 – Cluster of Excellence "Physics of Life" and EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of TU Dresden, and DFG grant 389792660 as part of TRR 248.

Images/graphs/plots/tables/data license/copyright

With the exception of those images from external authors whose licenses/copyrights we have specified in the respective figure captions, we as authors state that all of our own figures, graphs, plots, and data tables in this article (i. e., those not marked) are and remain under our own personal copyright, with the permission to be used here. We also make them available under the Creative Commons Attribution 4.0 International (🅭🅯 CC BY 4.0) license and share them at `osf.io/fxs76`.

REFERENCES

[1] K. Ajani, E. Lee, C. Xiong, C. N. Knaflic, W. Kemper, and S. Franconeri. Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Trans Vis Comput Graph*, 2022. To appear. doi: 10.1109/TVCG.2021.3068337

[2] B. Bach, P. Dragicevic, S. Huron, P. Isenberg, Y. Jansen, C. Perin, A. Spritzer, R. Vuillemot, W. Willett, and T. Isenberg. Illustrative data graphics in 18th–19th century style: A case study. In *Posters of IEEE VIS*, 2013. https://hal.inria.fr/hal-00849079.

[3] M. Balzer and O. Deussen. Voronoi treemaps. In *Proc. InfoVis*, pp. 49–56. IEEE Comp. Soc., Los Alamitos, 2005. doi: 10.1109/INFVIS.2005.1532128

[4] M. S. Bartlett. A note on the multiplying factors for various χ^2 approximations. *J R Stat Soc B*, 16(2):296–298, 1954.

[5] F. Beck, M. Burch, and S. Diehl. Towards an aesthetic dimensions framework for dynamic graph visualisations. In *Proc. IV*, pp. 592–597. IEEE Comp. Soc., Los Alamitos, 2009. doi: 10.1109/IV.2009.42

[6] B. Behrendt, P. Berg, O. Beuing, B. Preim, and S. Saalfeld. Explorative blood flow visualization using dynamic line filtering based on surface features. *Comput Graph Forum*, 37(3):183–194, June 2018. doi: 10.1111/cgf.13411

[7] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch. The aesthetics of graph visualization. In *Proc. CAe*. Eurographics Assoc., Goslar, 2007. doi: 10.2312/COMPAESTH/COMPAESTH07/057-064

[8] M. Bentvelzen, J. Niess, M. P. Woźniak, and P. W. Woźniak. The development and validation of the technology-supported reflection inventory. In *Proc. CHI*, pp. 366:1–366:8. ACM, New York, 2021. doi: 10.1145/3411764.3445673

[9] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. Esri Press, 1983.

[10] J. Bertin. *Sémiologie Graphique*. Éd. de l'EHESS, Paris, 3rd ed., 1998.

[11] L. Besançon, A. Semmo, D. Biau, B. Frachet, V. Pineau, E. H. Sariali, M. Soubeyrand, R. Taouachi, T. Isenberg, and P. Dragicevic. Reducing affective responses to surgical images and videos through stylization. *Comput Graph Forum*, 39(1):462–483, Feb. 2020. doi: 10.1111/cgf.13886

[12] J. Blijlevens, C. Thurgood, P. Hekkert, L.-L. Chen, H. Leder, and T. Whitfield. The aesthetic pleasure in design scale: The development of a scale to measure aesthetic pleasure for designed artifacts. *Psychol Aesthet Creat Arts*, 11(1):86–98, Feb. 2017. doi: 10.1037/aca0000098

[13] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front Public Health*, 6:149:1–149:18, June 2018. doi: 10.3389/fpubh.2018.00149

[14] A. Boomsma and J. J. Hoogland. The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, and D. Sörbom, eds., *Structural Equation Models: Present and Future. A Festschrift in Honor of Karl Jöreskog*, pp. 139–168. Scientific Software International, Lincolnwood, 2001.

[15] R. Brath, M. Peters, and R. Senior. Visualization for communication: The importance of aesthetic sizzle. In *Proc. IV*, pp. 724–729. IEEE Comp. Soc., Los Alamitos, 2005. doi: 10.1109/IV.2005.145

[16] A. Brielmann. Empirical aesthetics. In *Internet Encyclopedia of Philosophy*. Accessed: March 2022; https://iep.utm.edu/emp-aest/.

[17] T. Büring, J. Gerken, and H. Reiterer. User interaction with scatterplots on small screens – A comparative evaluation of geometric-semantic zoom and fisheye distortion. *IEEE Trans Vis Comput Graph*, 12(5):829–836, Sept./Oct. 2006. doi: 10.1109/TVCG.2006.187

[18] L. Byron and M. Wattenberg. Stacked graphs – Geometry & aesthetics. *IEEE Trans Vis Comput Graph*, 14(6):1245–1252, Nov./Dec. 2008. doi: 10.1109/TVCG.2008.166

[19] N. Cawthon and A. Vande Moere. A conceptual model for evaluating aesthetic effect within the user experience of information visualization. In *Proc. IV*, pp. 374–382. IEEE Comp. Soc., Los Alamitos, 2006. doi: 10.1109/IV.2006.4

[20] N. Cawthon and A. Vande Moere. The effect of aesthetic on the usability of data visualization. In *Proc. IV*, pp. 637–648. IEEE Comp. Soc., Los Alamitos, 2007. doi: 10.1109/IV.2007.147

[21] C. Chen. Top 10 unsolved information visualization problems. *IEEE Comput Graph Appl*, 25(4):12–16, July/Aug. 2005. doi: 10.1109/MCG.2005.91

[22] S. Chen, N. Andrienko, G. Andrienko, J. Li, and X. Yuan. Co-bridges: Pair-wise visual connection and comparison for multi-item data streams. *IEEE Trans Vis Comput Graph*, 27(2):1612–1622, Feb. 2020. doi: 10.1109/TVCG.2020.3030411

[23] D. Child. *The Essentials of Factor Analysis*. Continuum International Publishing Group, London, 3rd ed., 2006.

[24] C. Collins, G. Penn, and S. Carpendale. Bubble Sets: Revealing set relations with isocontours over existing visualizations. *IEEE Trans Vis Comput Graph*, 15(6):1009–1016, Nov./Dec. 2009. doi: 10.1109/TVCG.2009.122

[25] B. Cornelissen, D. Holten, A. Zaidman, L. Moonen, J. J. van Wijk, and A. van Deursen. Understanding execution traces using massive sequence and circular bundle views. In *Proc. ICPC*, pp. 49–58. IEEE Comp. Soc., Los Alamitos, 2007. doi: 10.1109/ICPC.2007.39

[26] R. F. DeVellis and C. T. Thorpe. *Scale Development: Theory and Applications*. Sage Publications, 5th ed., 2021.

[27] I. K. Duncan, S. Tingsheng, S. T. Perrault, and M. T. Gastner. Task-based effectiveness of interactive contiguous area cartograms. *IEEE Trans Vis Comput Graph*, 27(3):2136–2152, Mar. 2020. doi: 10.1109/TVCG.2020.3041745

[28] D. Dutton. *The Art Instinct: Beauty, Pleasure, and Human Evolution*. Oxford University Press, USA, 2009.

[29] L. R. Fabrigar and D. T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, 2012. doi: 10.1093/acprof:osobl/9780199734177.001.0001

[30] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods*, 4(3):272–299, Sept. 1999. doi: 10.1037/1082-989X.4.3.272

[31] J. A. Gliem and R. R. Gliem. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, pp. 82–88, 2003. https://hdl.handle.net/1805/344.

[32] L. K. M. Graf and J. R. Landwehr. A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Pers Social Psychol Rev*, 19(4):395–410, Nov. 2015. doi: 10.1177/1088868315574978

[33] L. K. M. Graf and J. R. Landwehr. Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Front Psychol*, 8:15:1–15:15, Jan. 2017. doi: 10.3389/fpsyg.2017.00015

[34] J. F. Hair. *Multivariate Data Analysis*. Pearson, 7th ed., 2009.

[35] L. Harrison, K. Reinecke, and R. Chang. Infographic aesthetics: Designing for the first impression. In *Proc. CHI*, pp. 1187–1190. ACM, New York, 2015. doi: 10.1145/2702123.2702545

[36] M. Hassenzahl, M. Burmester, and F. Koller. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer*, pp. 187–196. Vieweg+Teubner, Wiesbaden, 2003. doi: 10.1007/978-3-322-80058-9_19

[37] C. G. Healey and J. T. Enns. Perception and painting: A search for effective, engaging visualizations. *IEEE Comput Graph Appl*, 22(2):10–15, Mar./Apr. 2002. doi: 10.1109/38.988741

[38] A. Inselberg. Multidimensional detective. In *Proc. InfoVis*, pp. 100–107. IEEE Comp. Soc., Los Alamitos, 1997. doi: 10.1109/INFVIS.1997.636793

[39] B. Jenny, M. Heitzler, D. Singh, M. Farmakis-Serebryakova, J. C. Liu, and L. Hurni. Cartographic relief shading with neural networks. *IEEE Trans Vis Comput Graph*, 27(2):1225–1235, Feb. 2020. doi: 10.1109/TVCG.2020.3030456

[40] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, Mar. 1974. doi: 10.1007/BF02291575

[41] P. Kok, M. Baiker, E. A. Hendriks, F. H. Post, J. Dijkstra, C. W. Lowik, B. P. Lelieveldt, and C. P. Botha. Articulated planar reformation for change visualization in small animal imaging. *IEEE Trans Vis Comput Graph*, 16(6):1396–1404, Nov./Dec. 2010. doi: 10.1109/TVCG.2010.134

[42] K. Lai and S. B. Green. The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivar Behav Res*, 51(2–3):220–239, Mar. 2016. doi: 10.1080/00273171.2015.1134306

[43] A. Lau and A. Vande Moere. Towards a model of information aesthetics in information visualization. In *Proc. IV*, pp. 87–92. IEEE Comp. Soc., Los Alamitos, 2007. doi: 10.1109/IV.2007.114

[44] T. Lavie and N. Tractinsky. Assessing dimensions of perceived visual aesthetics of web sites. *Int J Hum Comput Stud*, 60(3):269–298, Mar. 2004. doi: 10.1016/j.ijhcs.2003.09.002

[45] H. Leder, B. Belke, A. Oeberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgments. *Br J Psychol*, 95(4):489–508, Nov. 2004. doi: 10.1348/0007126042369811

[46] R. A. Likert. A technique for the measurement of attitudes. *Arch Psychol*, 22(140):5–55, 1932.

[47] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. StoryFlow: Tracking the

evolution of stories. *IEEE Trans Vis Comput Graph*, 19(12):2436–2445, Dec. 2013. doi: 10.1109/TVCG.2013.196

[48] J. Mankoff, A. K. Dey, G. Hsieh, J. Kientz, S. Lederer, and M. Ames. Heuristic evaluation of ambient displays. In *Proc. CHI*, pp. 169–176. ACM, New York, 2003. doi: 10.1145/642611.642642

[49] G. E. Marai, B. Pinaud, K. Bühler, A. Lex, and J. H. Morris. Ten simple rules to create biological network figures for communication. *PLoS Comput Biol*, 15(9):e1007244:1–e1007244:16, Sept. 2019. doi: 10.1371/journal.pcbi.1007244

[50] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao. A global database of COVID-19 vaccinations. *Nat Hum Behav*, 5:947–953, July 2021. doi: 10.1038/s41562-021-01122-8

[51] M. Minge, M. Thüring, I. Wagner, and C. V. Kuhr. The meCUE questionnaire: A modular tool for measuring user experience. In *Advances in Ergonomics Modeling, Usability & Special Populations*, pp. 115–128. Springer, Cham, 2017. doi: 10.1007/978-3-319-41685-4_11

[52] M. Moshagen and M. T. Thielsch. Facets of visual aesthetics. *Int J Hum Comput Stud*, 68(10):689–709, Oct. 2010. doi: 10.1016/j.ijhcs.2010.05.006

[53] M. Nadal, A. Gomila, and A. Gàlvez-Pol. A history for neuroaesthetics. In J. O. Lauring, ed., *An Introduction to Neuroaesthetics: The Neuroscientific Approach to Aesthetic Experience, Artistic Creativity, and Arts Appreciation*. Museum Tusculanum Press, Univ. Copenhagen, 2013.

[54] M. Nadal and O. Vartanian. Empirical aesthetics: An overview. In M. Nadal and O. Vartanian, eds., *The Oxford Handbook of Empirical Aesthetics*. Oxford University Press, 2022. doi: 10.1093/oxfordhb/9780198824350.013.1

[55] Q. Nguyen, P. Eades, and S.-H. Hong. On the faithfulness of graph visualizations. In *Proc. GD*, pp. 566–568. Springer, Berlin, 2012. doi: 10.1007/978-3-642-36763-2_55

[56] C. Nobre, D. Wootton, L. Harrison, and A. Lex. Evaluating multivariate network visualization techniques using a validated design and crowdsourcing approach. In *Proc. CHI*, pp. 254:1–254:12. ACM, New York, 2020. doi: 10.1145/3313831.3376381

[57] J. C. Nunnally and I. H. Bernstein. *Psychometric Theory*. McGraw-Hill, 3rd ed., 1994.

[58] J. W. Osborne, A. B. Costello, and J. T. Kellow. Best practices in exploratory factor analysis. In J. W. Osborne, ed., *Best Practices in Quantitative Methods*, chap. 6, pp. 86–99. Sage, 2020. doi: 10.4135/9781412995627.d8

[59] H. C. Purchase. Metrics for graph drawing aesthetics. *J Vis Lang Comput*, 13(5):501–516, Oct. 2002. doi: 10.1006/jvlc.2002.0232

[60] R. Reber. Appreciation modes in empirical aesthetics. In M. Nadal and O. Vartanian, eds., *The Oxford Handbook of Empirical Aesthetics*. Oxford University Press, 2021. doi: 10.1093/oxfordhb/9780198824350.013.38

[61] R. Reber, N. Schwarz, and P. Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Pers Social Psychol Rev*, 8(4):364–382, Nov. 2004. doi: 10.1207/s15327957pspr0804_3

[62] W. Revelle. psych: Procedures for psychological, psychometric, and personality research. R package, 2022. https://CRAN.R-project.org/package=psych.

[63] Y. Rosseel. lavaan: An R package for structural equation modeling. *J Stat Software*, 48(2):1–36, 2012. doi: 10.18637/jss.v048.i02

[64] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods Psychol Res Online*, 8(2):23–74, 2003.

[65] M. Schrepp, J. Thomaschewski, and A. Hinderks. Construction of a benchmark for the user experience questionnaire (UEQ). *Int J Interact Multimedia Artif Intell*, 4(4):40–44, June 2017. doi: 10.9781/ijimai.2017.445

[66] T. Schultz and G. L. Kindlmann. Superquadric glyphs for symmetric second-order tensors. *IEEE Trans Vis Comput Graph*, 16(6):1595–1604, Sept./Dec. 2010. doi: 10.1109/TVCG.2010.199

[67] S. Stusak, A. Tabard, F. Sauka, R. A. Khot, and A. Butz. Activity sculptures: Exploring the impact of physical visualizations on running activity. *IEEE Trans Vis Comput Graph*, 20(12):2201–2210, Dec. 2014. doi: 10.1109/TVCG.2014.2352953

[68] L. G. Tateosian, C. G. Healey, and J. T. Enns. Engaging viewers through nonphotorealistic visualizations. In *Proc. NPAR*, pp. 93–102. ACM, New York, 2007. doi: 10.1145/1274871.1274886

[69] C. Tominski and H. Schumann. Enhanced interactive spiral display. In *Proc. SIGRAD*, pp. 53–56. Linköping University Electronic Press, Sweden, 2008.

[70] A. Vande Moere and H. Purchase. On the role of design in information visualization. *Inf Vis*, 10(4):356–371, Oct. 2011. doi: 10.1177/1473871611415996

[71] R. J. Vandenberg. Introduction: Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea? *Organ Res Methods*, 9(2):194–201, Apr. 2006. doi: 10.1177/1094428105285506

[72] I. Viola, M. Chen, and T. Isenberg. Visual abstraction. In M. Chen, H. Hauser, P. Rheingans, and G. Scheuermann, eds., *Foundations of Data Visualization*, chap. 2, pp. 15–37. Springer, Berlin, 2020. doi: 10.1007/978-3-030-34444-3_2

[73] I. Viola and T. Isenberg. Pondering the concept of abstraction in (illustrative) visualization. *IEEE Trans Vis Comput Graph*, 24(9):2573–2588, Sept. 2018. doi: 10.1109/TVCG.2017.2747545

[74] Y. Wang, A. Segal, R. Klatzky, D. F. Keefe, P. Isenberg, J. Hurtienne, E. Hornecker, T. Dwyer, and S. Barrass. An emotional response to the value of visualization. *IEEE Comput Graph Appl*, 39(5):8–17, Sept./Oct. 2019. doi: 10.1109/MCG.2019.2923483

[75] M. W. Watkins. Exploratory factor analysis: A guide to best practice. *J Black Psychol*, 44(3):219–246, Apr. 2018. doi: 10.1177/0095798418771807

[76] M. Weiß, K. Angerbauer, A. Voit, M. Schwarzl, M. Sedlmair, and S. Mayer. Revisited: Comparison of empirical methods to evaluate visualizations supporting crafting and assembly purposes. *IEEE Trans Vis Comput Graph*, 27(2):1204–1213, Feb. 2020. doi: 10.1109/TVCG.2020.3030400

[77] J. Woodring, M. Petersen, A. Schmeißer, J. Patchett, J. Ahrens, and H. Hagen. In situ eddy analysis in a high-resolution ocean climate model. *IEEE Trans Vis Comput Graph*, 22(1):857–866, Jan. 2016. doi: 10.1109/TVCG.2015.2467411

[78] K. Xu, C. Rooney, P. Passmore, D.-H. Ham, and P. H. Nguyen. A user study on curved edges in graph visualization. *IEEE Trans Vis Comput Graph*, 18(12):2449–2456, Dec. 2012. doi: 10.1109/TVCG.2012.189

[79] F. Yang-Wallentin and K. G. Jöreskog. Robust standard errors and Chi-squares for interaction models. In *New Developments and Techniques in Structural Equation Modeling*, chap. 6, pp. 159–171. Psychology Press, New York, 2001. doi: 10.4324/9781410601858-11