AMERICAN PSYCHOLOGICAL ASSOCIATION
DIVISION 10
Psychology of Aesthetics, Creativity, and the Arts

© 2020 American Psychological Association
ISSN: 1931-3896
http://dx.doi.org/10.1037/aca0000331

Why People Press "Like": A New Measure for Aesthetic Appeal Derived From Instagram Data

Katja Thömmes and Ronald Hübner
University of Konstanz

Be it on Facebook, Twitter, or Instagram, "Like" buttons are all over social media generating huge amounts of data. In this project, we develop methods for leveraging Instagram data with the purpose of developing a measure that is useful as a proxy for the aesthetic appeal of photographs. Based on the metadata of 15,073 photographs from the photographic genres of architecture, dance, and landscape gathered from 9 different Instagram accounts of professional photographers, we compute the Image Aesthetic Appeal score (IAA). We conduct an online experiment to test how IAA scores relate to more commonly used psychological variables, such as rating scales of aesthetic liking. We also investigate both low-level features and content-related preferences in the image set. Our results show that IAA scores are a reliable measure and are substantially predictive for experimental liking data. Also, solidly established preferences for curvature and visual balance as well as content-related preferences are reflected in IAA scores. We provide a simple method to compute IAA scores for all sorts of content on Instagram. This prepares ground for investigating large data sets with regard to aesthetic appeal of photographic images.

Keywords: measuring aesthetics, aesthetic appeal, social media, Instagram, photography

Why do we find something beautiful? What makes the aesthetic appeal of an image? In the field of experimental psychology, these questions are usually answered through experiments, in which participants are asked to rate images of artworks or other objects with respect to how much they like them. It is then examined to what extent one or more image features can account for the variance in liking ratings. However, for large image sets it is quite costly and time consuming to gather aesthetic ratings. Furthermore, the participants' task to rate oftentimes hundreds of visual stimuli in terms of liking in an experimental setting differs from how people evaluate what they like in real life, which might restrict the validity of experimental data. Therefore, in the present article we establish a new method to quantify the aesthetic appeal of images. Rather than collecting experimental data, we propose utilizing readily available online data, or, more precisely, online *liking* data from the social media platform Instagram. We believe that this approach has great potential, because it is an inexpensive and virtually limitless source of data that covers everyday online behavior of more than a billion Instagram users. It is not only promising for empirical research on aesthetics, but also for computational aesthetics. In the latter field, where algorithms are trained to automatically assess the aesthetic appeal of images, such large scale data is of high value for both training and validation (Siahaan, Hanjalic, & Redi, 2016).

Instagram "Likes" have already been investigated as a measure of aesthetic appeal for architectural photographs (Thömmes & Hübner, 2018), where there is a positive correlation of Likes with aesthetic preference choices collected in an experiment. Moreover, the number of Likes could partly be predicted by low-level features such as visual balance and curvature. However, the application of that method is rather restricted, because it does not allow to compare pictures that were uploaded far apart in time. The reason is that numbers of Likes usually increase together with growing numbers of followers over time. In the present study we solved this problem by developing a method to discount the follower effect. For this objective we analyzed data of about 15,000 images from three genres: architecture, dancer portraits, and landscapes. The new method was the prerequisite for defining an aesthetic score that allows to compare the aesthetic appeal of photographs within and across Instagram accounts.

Measuring Aesthetic Appeal

Reviewing the literature on measuring aesthetic appeal in the visual domain, we found a great variety in scales and a broad range of methods, including preference ranking tasks (Axelsson, 2007), photo quality assessment by experts (Cerosaletti & Loui, 2009) or

Katja Thömmes and Ronald Hübner, Cognitive Psychology, Department of Psychology, University of Konstanz.

We are grateful to the following photographers for granting permission to use their photographs and Instagram data for scientific purposes: @fernsehturm_, @le_blanc, @macenzo, @karolinakuras, @NYCdanceproject, @rachelnevillephoto, @janske, @jn, and @othellonine.

Correspondence concerning this article should be addressed to Katja Thömmes, Cognitive Psychology, Department of Psychology, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Baden-Württemberg. E-mail: katja.thoemmes@uni-konstanz.de

by crowdsourcing (Lebreton, Raake, & Barkowsky, 2016), liking scales (Gershoni & Kobayashi, 2006; Tinio, Leder, & Strasser, 2011), as well as rating scales for various aesthetic descriptors, including beautiful-ugly (Jacobsen, Buchta, Köhler, & Schröger, 2004), attractive-unattractive, pleasant-unpleasant, and interesting-uninteresting (Russell & George, 1990). There is also a growing body of research dealing with deep learning and convolutional neural networks (CNN) to rate aesthetics of visual stimuli (e.g., Kong, Shen, Lin, Mech, & Fowlkes, 2016) and to model aesthetic perception (e.g., Denzler, Rodner, & Simon, 2016). Despite a strong interest in understanding what makes images aesthetically appealing, there has been little research on the reliability and validity of aesthetic measures as such. To the best of our knowledge, Siahaan, Hanjalic, and Redi (2016) conducted the first systematic study of how different experimental environments (lab vs. crowdsourcing) and different rating scales influence the reliability and repeatability of aesthetic evaluations in the visual domain. They found that a discrete 5-point scale yielded the most reliable results compared to a continuous scale with and without visual anchors and to a binary scale—though all scales led to reliable results. Concerning the experimental environment, they concluded that aesthetic appeal evaluations are repeatable between lab and crowdsourcing environments, albeit reliability slightly decreased for all scales in the less controlled crowdsourcing experiments. They highlighted that an extension to even less controlled data collected from photo sharing websites needs further investigation. Our project is an attempt to fill this methodological gap.

Murray, Marchesotti, and Perronnin (2012) chose an approach similar to ours when setting up their database for Aesthetic Visual Analysis (AVA) that contains over 250,000 images and aesthetic quality scores gathered from the photo sharing platform dpchallenge.com. Their aesthetic scores are votes on the platform given by amateur and professional photographers on a 10-point-scale for image quality (images received an average of 200 votes). The AVA database is a valuable basis for investigating the interplay between photographic style and aesthetic quality and it is unique in its size. However, the authors did not provide any test of reliability or experimental validation of these aesthetic scores. Also, the aesthetic quality ratings stem from voters that show keen interest in photography and must be considered experts. In our view, *aesthetic quality* in that sense differs from the more general concept of *aesthetic appeal*. In a very literal sense, the aesthetic appeal of an image translates to the image's capacity to appeal to people in a way that they would call it beautiful, likable, interesting, or aesthetically pleasing (for a full discussion of the language of aesthetics in the visual domain see Augustin, Wagemans, & Carbon, 2012). We want to offer a methodology to compute and validate a measure of aesthetic appeal by investigating liking behavior of the much larger and much more diverse audience on Instagram, where images receive an average of many thousands of Likes and reach up to hundreds of thousands of viewers.

When taking an objectivist perspective on visual aesthetics, one aims at examining features of the object that account for its aesthetic appeal rather than the often-cited "eye of the beholder" or the aesthetic experience of individuals. It is assumed that there is a *true aesthetic appeal score* for every visual stimulus. This approach is as old as Fechner (1876), the founder of empirical aesthetics, who proposed that the aesthetic appeal of any visual stimulus can be measured and investigated psychophysically. To describe this aesthetic property, Fechner used the German phrase "das wahrhaft Schöne" that is best translated as "that which is truly beautiful." Fechner supposed that this inherent beauty of any object defines how much it appeals to people[1] (Fechner, 1876, p. 16). However, as Berlyne states in his seminal book *Aesthetics and Psychobiology* (Berlyne, 1971, p. 7), we have to rely on "the scientific study of *aesthetic behavior* of the *appreciator,"* in order to measure the aesthetic appeal of an image.[2] Whereas Berlyne's work on aesthetics focuses on artworks, we find it reasonable to generalize his idea to other types of visual stimuli and to photography in particular. Hence, in the Instagram context, this translates to *users* (appreciators), who spend their time looking at images and *pressing the Like button* (aesthetic behavior), if an image appeals to them. We aim at transforming this aesthetic behavior of millions of people on Instagram into an aesthetic score that could potentially be calculated for every single photo on the platform and then be used to investigate what makes "good" image composition.

It might be argued that aesthetic preferences differ across photographic genres and that not all individuals, naturally, agree on whether one image is aesthetically more appealing than another one. Agreement among individuals largely depends on the image type. Vessel, Stahl, Maurer, Denker, and Starr (2014), for instance, found that people show higher levels of agreement when rating the aesthetic appeal of faces or landscapes compared to images of architecture or artworks. The authors interpreted their findings on architecture and artworks—low agreement among different people coupled with strong reliability for individual observers—"as evidence that preferences are not universally determined by specific visual features [of the object], but rather on the basis of subjective associations" (Vessel et al., 2014, p. 6). We want to propose a different perspective on shared taste or the lack of such. When looking at any visual stimulus, there are three aspects contributing to its aesthetic appeal: (a) context, (b) content, and (c) composition (Thömmes & Hübner, 2018). Low levels of agreement are likely caused by aspects of content. Why there are higher levels of shared taste for some genres compared to others can and should certainly be studied (Vessel, Maurer, Denker, & Starr, 2018). At the same time, nonetheless, there are effects of low-level features of the composition even in low agreement genres as was shown for architectural photographs (Thömmes & Hübner, 2018). Those effects are arguably a much smaller piece of the aesthetic puzzle. However, we are convinced that they are more universal than content and context variables. The latter often overshadow small but relevant effects of low-level features (Matz, Gladstone, & Stillwell, 2017) in complex stimulus material. In addition, from a methodological point of view, low agreement means high variance in aesthetic evaluations on rating scales by different individuals for single images. Averaging such ratings will result in overall less variance in the liking variable compared with image sets where people agree on whether they like single images or dislike them. This implies a range restriction of average liking ratings for low-agreement genres. When investigating the effects of low-level

[1] "[Der Begriff] des Schönen in einem engsten Sinne, des wahrhaft Schönen, des ächten Schönen, was nicht blos aus höherm Gesichtspuncte gefällt, sondern auch Recht hat zu gefallen [. . .]" (Fechner, 1876, p. 16).

[2] "as far as aesthetics is concerned, the experimental psychologist or psychobiologist must concentrate on the scientific study of aesthetic behaviour" (Berlyne, 1971, p. 7).

features on liking in such genres, effect sizes therefore are potentially underestimated (Bobko, Roth, & Bobko, 2001). All of these issues can be addressed when using Instagram data: First, very large data sets make the detection of small effects possible, allowing us to generalize findings from simple stimuli to more complex photographs. Second, there is no need for using averages. Measuring aesthetic appeal with Instagram data is based on the number of people who pressed "Like" and thus found the image appealing. The distribution of these Like counts should not depend on shared taste as long as there is no "Dislike" button.

Instagram

In this article we introduce and validate a measure of the aesthetic appeal of Instagram photos. The measure is called the Image Aesthetic Appeal (IAA) score and is based on data from Instagram. In the pursuit of this objective a number of confounds have to be addressed.

The social media platform Instagram reached one billion monthly active users in June 2018 after passing the 800 million user mark only 9 months earlier (Systrom, 2018). Similar to other social media platforms (e.g., Facebook, YouTube, Twitter), Instagram offers its users the possibility to upload, create and interact with content. Back in 2016, Instagram disclosed that its users upload 95 million photos and generate an average of 4.2 billion Likes on a daily basis (Abutaleb, 2016). At that time Instagram had only around 500 million users, so the numbers are likely much higher today. For a wide variety of academic disciplines, the data generated and collected by social media platforms have become a point of interest. Lee (2016) coined the term "Likeology" and his tutorial[3] at the ACM WebSci, 2016 offers a wide-ranging overview of research on online liking, including Like buttons (Facebook, Instagram), thumbs-up (YouTube), +1 buttons (Google+), Favorites (Twitter, Flickr), Upvotes (Reddit), Re-Pins (Pinterest), and star ratings (Amazon). Ferrara, Interdonato, and Tagarelli (2014) conducted a comprehensive study of the Instagram system suggesting that it can be considered as proxy of the real world and used to investigate human behavior at scale.

When opening the Instagram app, users first see their personalized Instagram Feed, where all posts of the accounts they follow appear. They can switch to the Instagram Explore section, where Instagram suggests content based on previous interactions, and where content can be searched based on keywords (i.e., hashtags). For a study of networks formed by follow and like activity on Instagram see also Jang, Han, and Lee (2015). Figure 1 shows what an Instagram post looks like. As of today (September 2019), there are four buttons below every post. Users can interact with the post by clicking the heart symbol (i.e., Like), write a comment by clicking the speech bubble, share the post with friends by clicking the paper plane symbol, or save the post with the flag symbol on the right. "Liking" is by far the most frequently used type of interaction (Ferrara et al., 2014).

There is some research dealing with motives behind clicking Like buttons on social media. Gan (2017) investigated users' liking behavior on WeChat (a Chinese social media app) and found what she calls *hedonic gratification* (enjoyment, free time activity) to be the most important factor that motivates online liking, along with social support, information seeking, and self-presentation. A general overview on online liking behavior on different social media platforms concluded that liking indicates enjoyment and appreciation of content (Lowe-Calverley & Grieve, 2018), which supports our idea of linking Likes to the aesthetic appeal of the images.

Instagram users and content are highly diverse. Users include private persons sharing snapshots of their daily lives, businesses promoting their latest products or services, so-called influencers and bloggers sharing their lifestyles, artists (e.g., photographers, painters, poets) posting their professional work, and many more. For researchers dealing with Instagram data, it is important to be aware of this diversity and to clearly define what type of content is to be explored. In the present study, we are interested in fundamentals of good photographic image composition. Therefore, we restricted our analysis to Instagram accounts of professional photographers working in one of three genres: architecture, dancer portraits, and landscape.

Leveraging Instagram Data

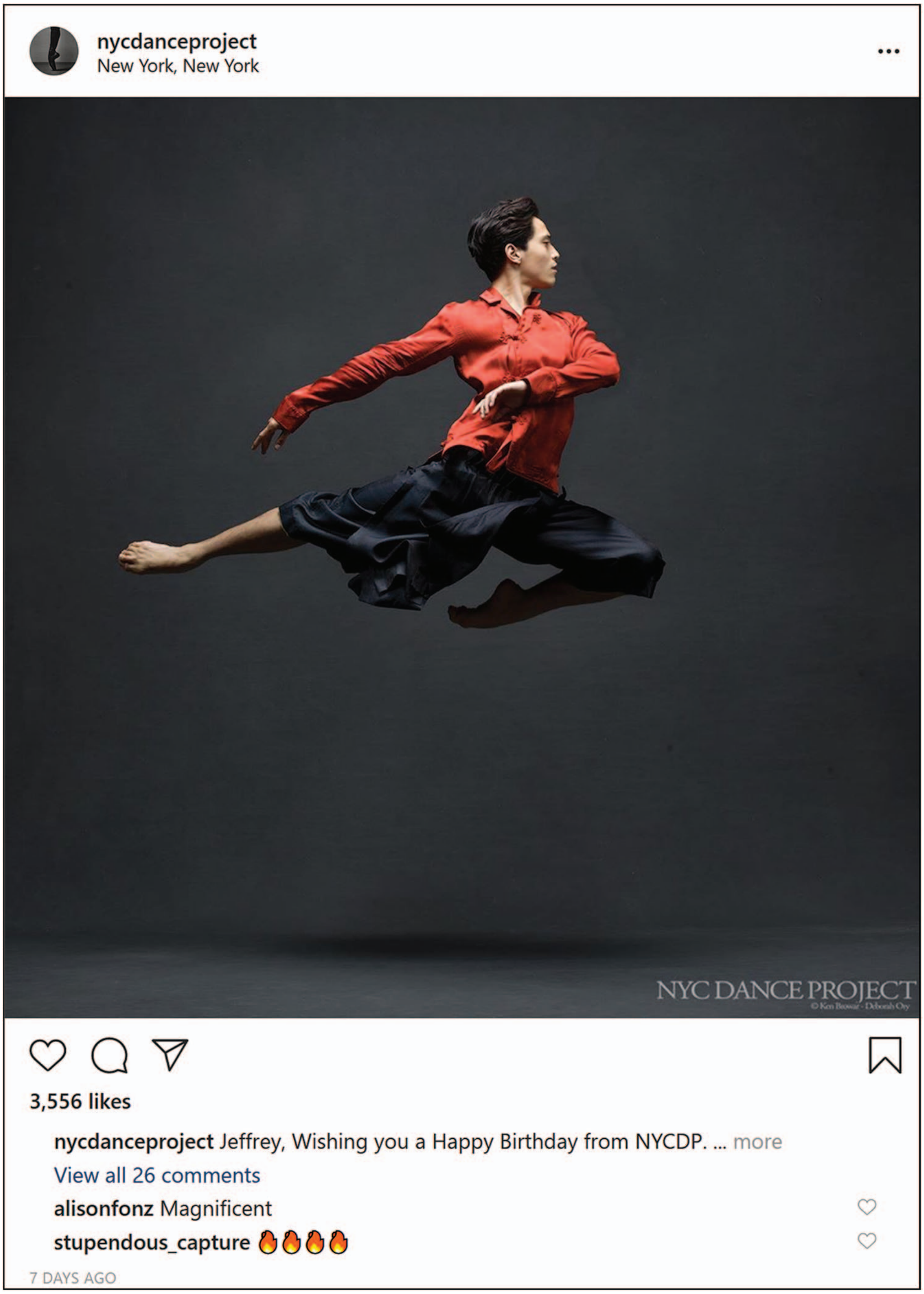
For our database, we chose well-established Instagram accounts that meet three inclusion criteria: First, as described above, we confined our analysis to Instagram accounts of professional photographers sharing mainly high-quality content. Second, the accounts must create homogenous content that classifies as either architecture, dancer portraits, or landscape photography. Consistent content is considered a basic rule for managing a successful Instagram account (Carroll, 2017). Jang et al. (2015) also found empirical evidence that specialists on Instagram receive five times more Likes from the community than generalists who post a mix of content. And third, we included accounts with at least 15,000 followers, as a large number of followers guarantees that a wide range of users sees and potentially likes the images. We started our analyses with nine different accounts for which we collected both account data and individual image data in April 2018. For data collection we used the service provided by minter.io and for image download that of 4Kstogram. See Appendix A for descriptive statistics and more details of the data set.

Utilizing Instagram data is of great interest in many areas. For instance, in social media marketing it is important to assess the quality and effect of marketing efforts (Komok, 2018a, 2018b). A widely used analytic measure in this respect is the so-called *engagement rate* (ER_i), which reflects the percentage of followers who interacted with a posted image i. Likes and comments are usually added up for computing the ratio. There are different ways of calculating engagement rates (Komok, 2018a). The basic formula is:

$$ER_i = \frac{Likes_i + Comments_i}{Followers_i} 100 \tag{1}$$

The idea is that a higher engagement rates indicate better quality content. Mean engagement rates for accounts are widely used as a key performance indicator and are the industry standard to assess the quality of social media marketing (Komok, 2018a). To evaluate this measure within our data set, we calculated the engagement rates for all images and the *mean* engagement rate for each of the nine accounts. An interesting finding is that there is a strong negative correlation between the *mean* engagement rates and numbers of followers across the

[3] http://pike.psu.edu/publications/websci16/.



nycdanceproject
New York, New York

3,556 likes

nycdanceproject Jeffrey, Wishing you a Happy Birthday from NYCDP. ... more

View all 26 comments

alisonfonz Magnificent

stupendous_capture 🔥🔥🔥🔥

7 DAYS AGO

Figure 1. This screenshot illustrates how posts appear in users' Instagram feeds. Screenshot of an Instagram post. Photo reprinted from Instagram, by nycdanceproject. K. Browar & D. Ory, 2020, retrieved from instagram.com/p/CAVz3-LB-IP/ .Copyright 2020 by Ken Browar & Deborah Ory. Reprinted with permission. See the online article for the color version of this figure.

nine accounts, $r = -.740$, $p = .023$. Such negative relation has already been observed by Komok (2018b): Engagement rates are lower for larger accounts. This demonstrates that the percentage of actively engaged followers decreases with a growing followership. Thus, it seems that the more followers an account has, the higher the percentage of *passive* followers who do not actually interact with content, which largely restricts the usefulness of engagement rates. Also, within accounts there is a significant decrease in engagement rates from early to later posts for all but two accounts (A1: $r = -.234^{**}$; A2: $r = -.449^{**}$; A3: $r = -.290^{**}$; D1: $r = -.393^{**}$; D2: $r = -.444^{**}$; D3: $r = -.066^{*}$; L1: $r = -.528^{**}$; L2: $r = .067^{**}$;

L3: $r = .047$ *ns*). It seems that engagement (Likes plus comments) grows slower than follower counts.

Changing numbers of followers are a great challenge for any kind of quality assessment score derived from Instagram data. For instance, engagement rates do not allow to draw comparisons between images from different accounts that differ in follower counts. Even within accounts, images posted at different points in time cannot be compared. Aside from follower effects, the sheer duration that an image has been online might affect its number of Likes. To investigate these issues, we examined our Instagram data closely to uncover how the number of followers and Likes increase over time both within and across different accounts.

Time Effects on Likes

It seems logical that the number of Likes an image receives increases over time as more people have the chance to see it and click the Like button. This might be true on platforms like Flickr, where images are presented based on content-related galleries and staff picks. However, timeliness plays a crucial role for the Instagram algorithm that decides what users see in their feed. Newer posts are always prioritized (Instagram Press, 2018). This suggests that images receive most Likes right after they were posted. To test this, we randomly selected five newly posted images from five different accounts and analyzed how the numbers of Likes increased as a function of time, with an observation period of 6 months.

Figure 2 shows how the numbers of Likes increase as a function of time after upload for the five example images. The pattern is similar across images and accounts. Almost all liking happens during the first three days, marked as the liking phase in Figure 2. After that initial phase the numbers of Likes hardly increase any further. We checked back after 3 months and 6 months, respectively, and found only minimal changes in Like counts. Comparing numbers of Likes after 7 days with those after six months shows that the numbers increased by only 1.79% to 3.63% for the five images (1.79%, 2.10%, 2.94%, 3.49%, 3.63%). This confirms that the liking phase is relatively short. Almost all liking for an image happens during the first days after posting.

Next, we considered the relation between time and Likes within single accounts taking a cross-sectional perspective. Figure 3 illustrates full liking data for all images ever posted on the nine accounts, gathered in March 2018 for all images that were online for more than a week. It becomes clear, that there is a strong positive correlation between time of upload and the number of Likes: The later an image is posted, the more Likes it receives. Considering that the numbers of followers are also higher in later points in time (second *y*-axis in Figure 3 that will be discussed in more detail below), it becomes clear that this relation is highly confounded with follower growth. We will now take a closer look at follower effects on Likes.

Discounting the Follower Effect on Likes

As mentioned, the relation between numbers of Likes, time, and numbers of followers is complex but relevant. The phase in which a posted image receives most of its Likes is rather short and confined to the first week after it is uploaded. Consequently, Likes should depend on the followers at that time, that is, on the historical number of followers. It might be expected that these two numbers are related only loosely, because Instagram allows all users to like and comment images regardless of whether they are followers or not.[4] There is no data available on how many people in fact saw the image (image *reach* counts are only visible to business accounts tracking these data). However, as Figure 3 illustrates, the number of followers strongly determines the number of Likes and appears to proficiently approximate the reach of an image. Thus, for the purpose of assessing the aesthetic appeal of a picture, this effect has to be discounted. To do so for a specific picture, however, we need to estimate the historical number of followers, which is usually not publicly available (Anna, 2018).

To solve this problem, we developed a formula that allows to estimate the historical number of followers for each picture. Empirical evidence suggests that follower growth mainly depends on posting activity. Accounts that do not post, do not grow. Anecdotal evidence suggests that people are most likely to click the "follow" button, right after they have seen freshly posted content. To estimate historical follower data, we therefore assume a linear follower growth per image and calculate the follower estimate F_i‘ for an image with posting index i as

$$F_i' = \frac{N_{fol}}{N_{img}} i \tag{2}$$

where N_{fol} denotes the current number of followers and N_{img} the current number of images which can be read off the Instagram account page. With this formula we calculated the historical number of followers F_i‘ for each picture separately for all nine accounts. Figure 3 illustrates the resulting data and also depicts data points for "real" historical follower counts that we gathered from minter.io in March 2018, when this was still possible. These data are quite fragmentary. However, they confirm that our estimations fit the real time course of follower growth sufficiently well.

The question now is to what extent the number of Likes can be accounted for by the historical number of followers. When investigating Likes as a function of historical followers (see Appendix B for graphs), it becomes clear that there is a strong relation that seems to be nonlinear for most accounts. Therefore, as possible regression models we tested a second-degree polynomial (quadratic) model and a logarithmic model to predict Likes L':

$$L'_{quadratic} = a + bF_i' + cF_i'^2 \tag{3}$$

$$L'_{\log} = a + b\log(F_i') \tag{4}$$

We found that the quadratic regression was superior. For the nine accounts the logarithmic model explained on average 49.32% of the variance, ranging from 38.1% (D2) to 54.5% (A2 and A3), whereas the quadratic model explained on average 61.16% of the variance, ranging from 41.7% (D2) to 71.8% (A3). Table 1 summarizes these results. For the purpose of optimally discounting the follower effect on Likes, we chose the quadratic model.

Thus, numbers of Likes can largely be predicted by the historical number of followers alone. Based on these functional relationships, we discount the follower effect and compute an aesthetic appeal measure (IAA) that we want to use as a proxy for the aesthetic appeal of the images.

[4] Despite users who use the "private" setting and share their content exclusively with followers.

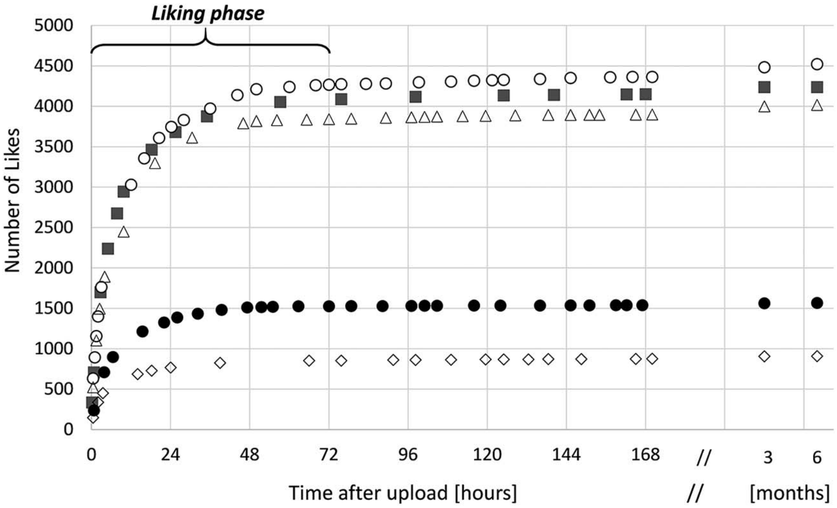


Figure 2. Numbers of Likes as a function of time after images were posted. We observed five images posted on five different accounts in April 2018 over a period of 6 months. Numbers of Likes remain stable after a relatively short initial liking phase of several days.

To calculate IAA scores that are independent of follower counts, we first calculated the follower predicted Likes *L'* (quadratic) for all images, separately for each of the nine accounts under investigation. Table 1 shows the equation coefficients *a*, *b*, and *c* for all accounts. We then calculated the absolute difference between predicted and observed Likes for each image. As can be seen in the graphs in Appendix B, the absolute prediction error increases with numbers of followers as the total variance increases. To correct for this increase, we calculated the percentage that the observed Likes L_i of image *i* deviate from the predicted Likes L_i' of that image, i.e.:

$$IAA_i = \frac{L_i - L_i'}{L_i'}100 \tag{5}$$

The result is a measure that is positive for images that receive more Likes than predicted by the number of followers alone, and negative for images that receive less Likes than predicted. As Figure 4 illustrates for account D2, this computation results in an evenly distributed score that does neither depend on followers, nor does it increase over time. Before drawing a small random sample of the images and compare the IAA scores to liking ratings from a small experiment, we chose two example images for demonstration purposes, one posted in 2014 when the account was still small, and another one posted late in 2017. Intuitive reasoning made us expect both a female dancer (vs. male) and color (vs. gray level) to increase the aesthetic appeal of the older image. As can be seen, the absolute Like counts do not reflect this preference, they even indicate the opposite (1,453 vs. 3,205 Likes). IAA scores, on the other hand, make the expected difference visible. The colorful photo with a female dancer scores at +23%, while the gray-level photo of a not even dancing male scores at −40%.

Figure 5 illustrates scatterplots of IAA scores for all nine accounts. For some accounts, IAA scores show a distinct starting growth (as do absolute Likes). Additionally, the initial photos were in some cases not related to the later theme, rather showing personal snapshots. Therefore, we consider it necessary to exclude these first images to prevent them from impairing the data. Due to individual differences between different accounts, we set a cut off by looking at each account separately. For account A2 we excluded the first 200 images, for D1 the initial 250 images, for D3 the initial 300 images, for L2 the initial 1,500 images, and for L3 the initial 200 images (see Appendix B). After excluding these images, IAA scores scatter evenly. When considering absolute Like counts there are some extreme outliers on the high end resulting in IAA scores of up to +500%. We checked some of these outliers and found external reasons for extreme liking behaviors in most cases. For example, there were features in photo magazines, or awards which the photographer won at the time of that specific post. Hence, extreme likings are not primarily motivated by the images' aesthetic appeal, but by external factors (context). We therefore decided to exclude extreme outliers and confine our analyses to IAA scores that lie within the range of −100% to +100%. It is important to set this cut off for IAA scores rather than absolute Likes, because IAA scores enable the detection of outlier images also in the early days when absolute levels of Likes were lower than later on. As a result, another 150 outlier images are excluded that lie above or below IAA scores of ±100.

Table 2 provides descriptive statistics for IAA scores for all nine accounts. The distribution of IAA scores is slightly positively skewed toward the high end, with more images receiving less Likes than followers alone predict (median of IAA scores <0 for all accounts) and more broadly dispersed larger scores. Pearson correlations of IAA cores with both the time variable and follower counts were calculated for all accounts. There were no systematic relationships across accounts. For some accounts the relation is slightly positive, for others slightly negative with low correlation coefficients ranging from −.185 to .336. We therefore concluded that IAA scores do not systematically depend on followers or time. With the remaining database containing 12,473 images, we inves-

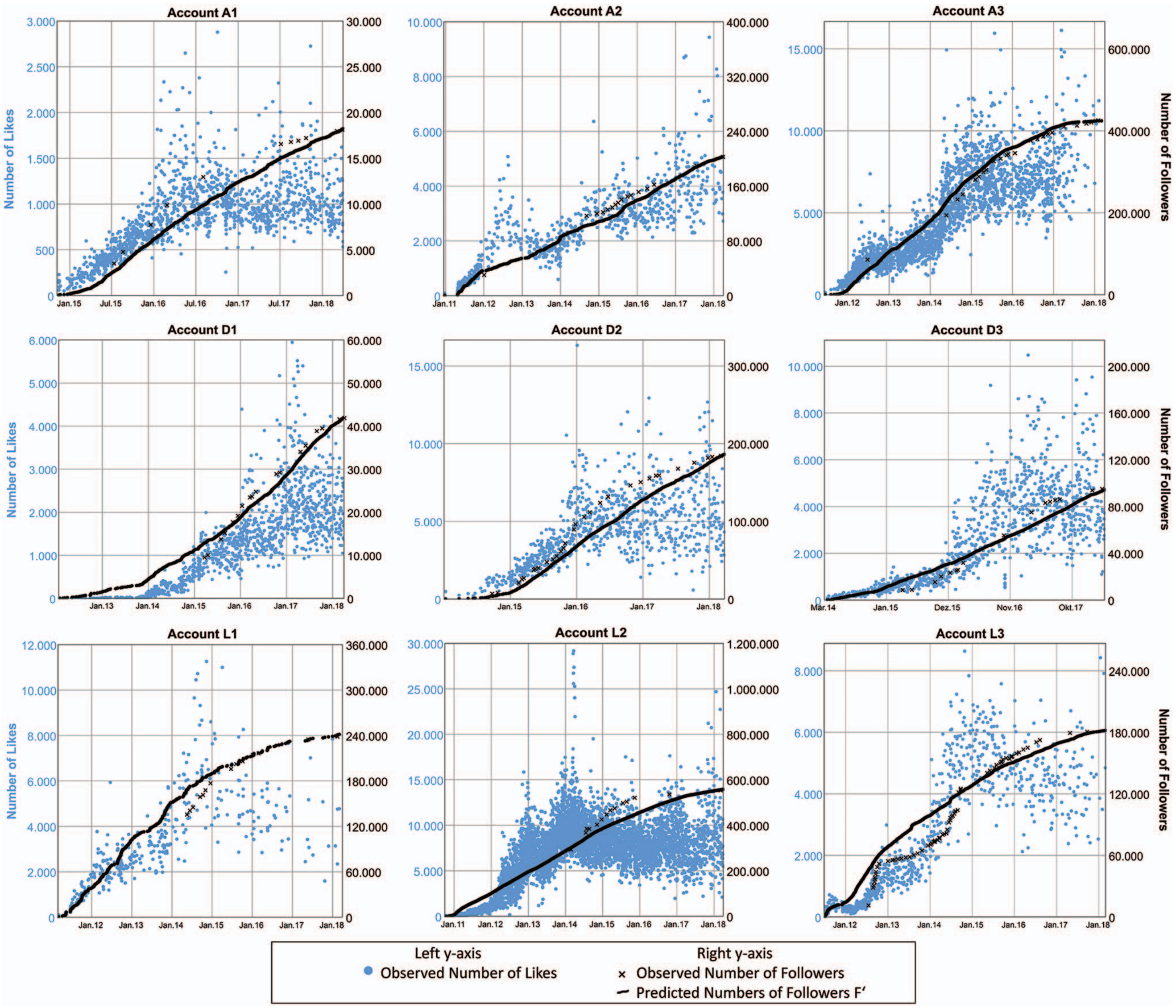


Figure 3. The blue/gray dots indicate numbers of Likes for each image (left *y*-axis). The black lines indicate corresponding historical follower counts that were estimated using formula (2). The black crosses indicate historical follower counts that we derived from minter.io (such data is no longer available since April 2018). See the online article for the color version of this figure.

tigated reliability and validity of IAA scores as a measure for aesthetic appeal.

Validating IAA Scores

Having developed the IAA score, questions concerning its reliability and validity arose. Do images receive the same score when posted again to another audience? Is the score similar to ratings obtained in an experiment despite its uncontrolled origin and many possible confounding variables on Instagram? And, finally, are IAA scores useful to investigate aesthetic fundamentals? To answer the first question, we selected images that were posted more than once on the same account. For the second question, we used IAA scores to predict ratings of aesthetic liking collected for a subset of the database in an online experiment. Concerning the third question, we investigated the low-level features curvature (Bar & Neta, 2006; Gómez-Puerto, Munar, & Nadal, 2016), and visual balance (Hübner & Fillinger, 2016; McManus, Stöver, & Kim, 2011; Wilson & Chatterjee, 2005), which are known to affect aesthetic preference, as well as one high-level factor: the gender of a portrayed dancer.

Reliability

Method

Analyzing the reliability of a measure developed from field data tends to be rather difficult. Fortunately, however, one of the

Table 1

R squares for Logarithmic and Quadratic Regressions Predicting Numbers of Likes With Followers (F_i') as Predictor Variable

			Parameter for quadratic regression equation used to calculate L'$_{\text{quadratic}}$ (unstandardized coefficients)		
Account[a]	Logarithmic model R^2	Quadratic model R^2	a	b	c
A1	.449**	.510**	130.02	0.152	-5.84×10^{-6}
A2	.545**	.642**	407.57	0.025	-2.24×10^{-8}
A3	.545**	.718**	355.34	0.023	-8.19×10^{-9}
D1	.518**	.649**	−463.08	0.131	-1.55×10^{-6}
D2	.381**	.417**	759.42	0.074	-2.73×10^{-7}
D3	.475**	.603**	−770.29	0.113	-6.07×10^{-7}
L1	.507**	.600**	−110.71	0.042	-7.99×10^{-8}
L2	.515**	.649**	−2322.36	0.057	-7.15×10^{-8}
L3	.504**	.716**	−750.40	0.042	-3.86×10^{-8}
Total	.552**	.688**			

Note. All p values for R^2 reach significance at ** $p < .01$.

[a] Three accounts per genre: architecture A1–A3, dancer D1–D3, landscape L1–L3. Instagram names of the accounts can be found in Appendix A.

accounts (D2) posted several images more than once, oftentimes even with the exact same captions and exactly one year apart. This enabled us to estimate test-retest-reliability for the IAA measure as well as engagement rates (see above) using intraclass correlation coefficients (McGraw & Wong, 1996). We calculated single measures intraclass correlation coefficients (ICCs) for 73 images, that were posted repeatedly (mean distance between posts was 355 days, *Min* = 47, *Max* = 972, *SD* = 228).

Results

The ICC estimates and their 95% confident intervals for both IAA scores and engagement rates were calculated using SPSS statistical package Version 25 based on a one-way random effects model (Koo & Li, 2016). The single measure ICC for the IAA measure was 0.770 with a 95% confidence interval from 0.658 to 0.849. The single measure ICC for engagement rates was 0.312 with a 95% confidence interval from 0.091 to 0.504. This indicates good test–retest reliability for the IAA scores and poor reliability for engagement rates according to the guidelines for interpreting reliability with ICCs by Koo and Li (2016).

Discussion

These results provide evidence that IAA scores are a reliable measure that does not depend on the passing of time or the

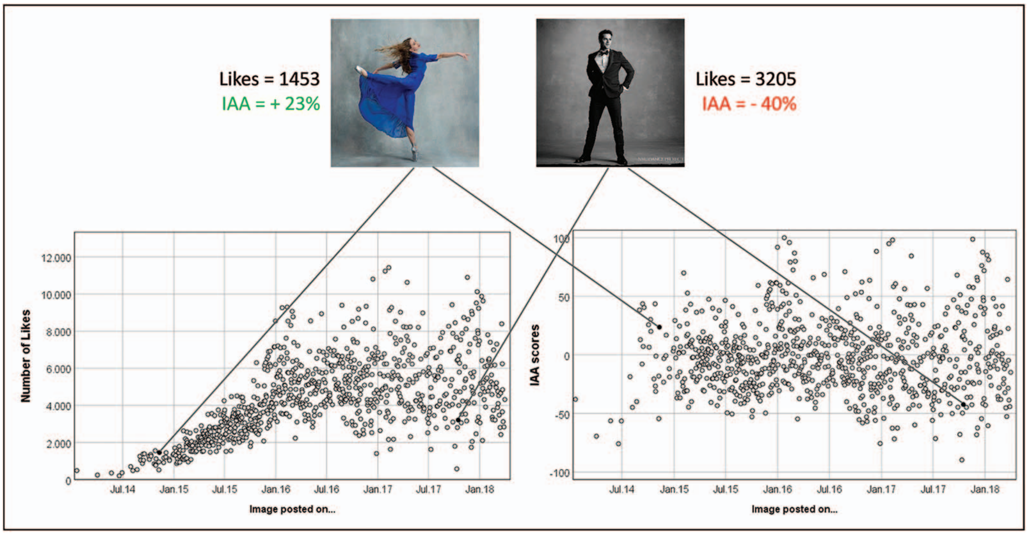


Figure 4. Each data point is an image posted by @NYCdanceproject. Absolute Likes (on the left) grow over time due to growing followership. IAA scores (on the right) are percentages above or below the amount of Likes that is explained by followers alone by polynomial regression. The two example images illustrate how IAA scores allow for better comparability of images within accounts. Absolute Likes vs. normalized Likes (IAA scores). Photos reprinted from Instagram, by nycdanceproject. K. Browar & D. Ory, 2020, retrieved from instagram.com/p/BaSPHS-FbAM/ & instagram.com/p/vMsGWgqmE-/. Copyright 2020 by Ken Browar & Deborah Ory. Adapted with permission. See the online article for the color version of this figure.

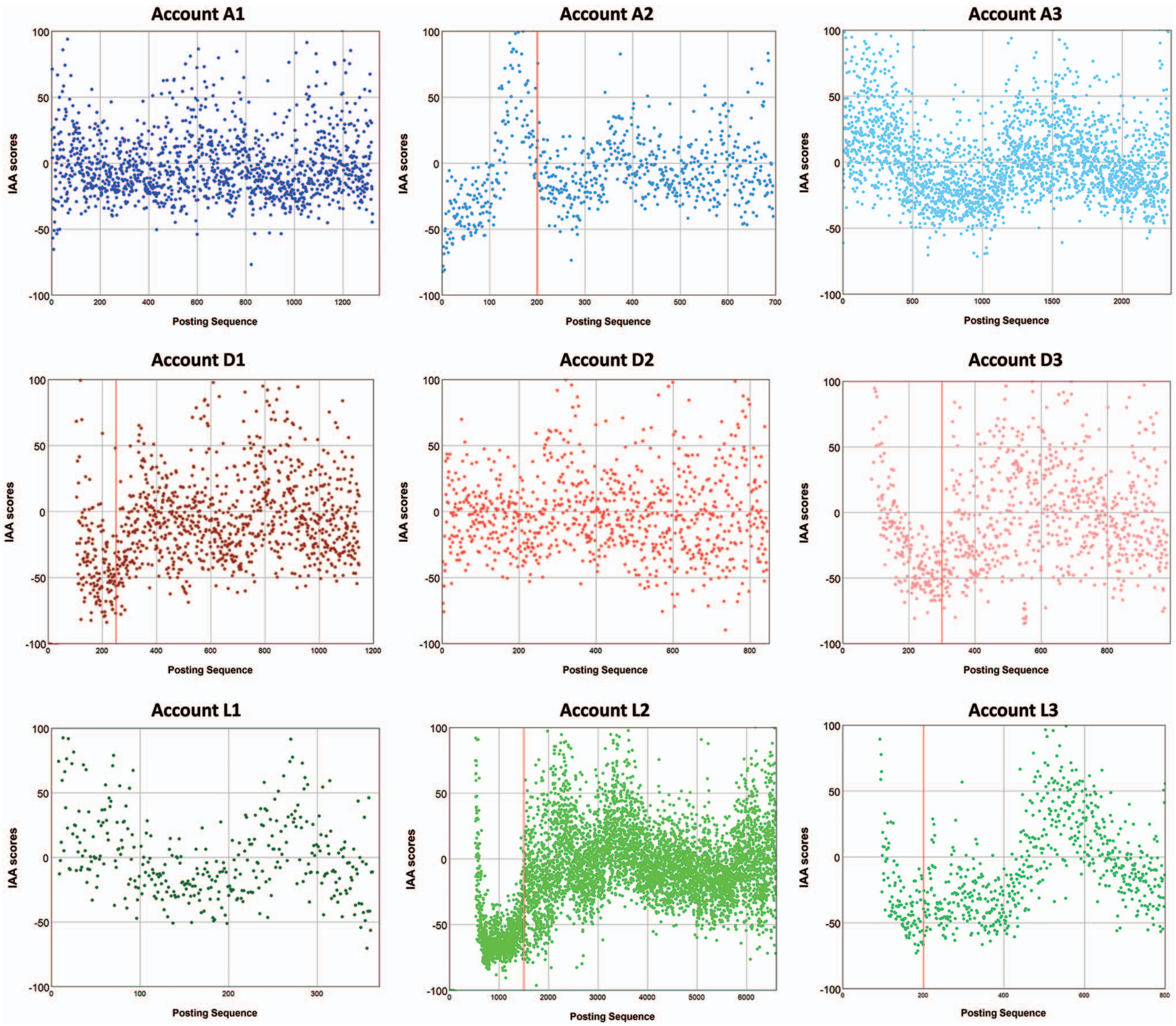


Figure 5. IAA scores for all images posted on all nine accounts as a function of posting sequence, after excluding scores below −100 and above +100. Red lines indicate the cut offs for early photos. See text for more information. See the online article for the color version of this figure.

correspondingly growing numbers of followers. The poor reliability of engagement rates is not surprising, given their dependence on followers as discussed above. However, the image sample we took for this analysis is rather small and also allows conclusions only about one account. Nevertheless, in view of the uncontrolled nature of Instagram data, these numbers made us optimistic and provided the basis for a further validation attempt of IAA scores to see whether they are an appropriate measure of aesthetic appeal.

Experimental Validation

To validate IAA scores as a proxy for aesthetic appeal, we conducted an online experiment in order to compare IAA scores to experimental liking data. Participants viewed and rated photographs from all three genres that appeared in a random order to approximate the Instagram experience of seeing different sorts of images. The study was performed in accordance with the ethical standards of the Declaration of Helsinki (1964) and its later amendments and with the ethics and safety guidelines of the University of Konstanz. Participants were informed that they are free to withdraw from the study at any point without any negative consequences. Informed consent was obtained from all participants by check-marking a box on the informed-consent page before the actual experiment started.

Method

We randomly selected 30 images from every account, resulting in 90 images per genre and 270 images in total. As is common in

Table 2
Descriptive Statistics for IAA Scores After Excluding Outliers and Starting Slope

Account[a]	Excluded outliers	Excluded starting slope	*N*	IAA *Median*	IAA *Mean*	IAA *SD*	IAA *Range*	Pearson *r* (IAA, time)	Pearson *r* (IAA, followers)
A1	17	0	1,304	−6.501	−2.025	24.44	170.75	−.003	.0001
A2	5	200	490	−8.268	−5.090	22.81	156.94	.157**	.150**
A3	23	0	2,306	−3.143	0.137	28.45	170.94	−.138**	−.124**
D1	17	250	880	−8.679	−4.766	30.98	175.57	.122**	.111**
D2	13	0	831	−3.283	−1.764	30.45	189.62	.013	.007
D3	13	300	667	−6.844	−2.960	36.51	184.61	.073	.067
L1	15	0	347	−5.165	−0.448	30.55	162.88	−.185**	−.182**
L2	45	1,500	5,050	−5.760	−3.099	26.80	196.05	−.017	−.024
L3	2	200	598	−13.015	−7.374	33.99	165.82	.263**	.336**
Total	150	2,450	12,473	−5.858	−2.619	28.43	196.26	−.018	−.017

** Pearson correlation is significant at the 0.01 Level (2-tailed).
[a] Three accounts per genre: architecture A1–A3, dancer D1–D3, landscape L1–L3. Instagram names of the accounts can be found in Appendix A.

experiments on aesthetics, the task of our participants was to indicate liking, that is, how much they liked a photograph on a visual analogue scale ranging from 0 (*I don't like it at all*) to 100 (*I like it very much*). In addition to rating the 270 images, each participant assessed 32 randomly selected images a second time to assess within-observer reliability. The online survey lasted about 20 min and was remunerated with 3 Euro.

Using Amazon's crowdsourcing platform Mechanical Turk (MTurk) for data collection, we recruited 80 participants. The participants used their own devices to participate, however, we asked them not to use tablets or smartphones. We gave explicit instructions on how to calibrate their screens: After the application automatically switched to full screen mode, participants were instructed to adjust the size of a standard page by using "CTRL +" or "CRTL −" to zoom in or out, so they would see the whole page as large as possible without having to scroll down on their screens.

We had to exclude seven of the 80 participants due to low within-observer reliability (< .50). The remaining 73 participants (36 females) were, on average, 38-years-old (age range: 22–77 years). Their mean within-observer reliability calculated on the basis of 32 repeated images was .86 (ranging from .59 to .99). We analyzed the data both in total and separately for the three genres. Table 3 sums up descriptive statistics, including *Means* and *SDs* for both MTurk liking and IAA scores.

Results

Pearson correlations of MTurk liking and IAA scores indicate a significant positive relation both across genres, $r = .347$, $p < .001$ and within genres (A: $r = .277$; D: $r = .575$; L: $r = .477$, always $p < .001$), see Table 3. When we control the relationship between MTurk liking and IAA scores for the genre variable we find an increased partial correlation of $r = .401$, $p < .001$. In addition, Table 4 sums up these results and shows Pearson correlations within all nine accounts. For two accounts correlations are non-significant, other two accounts have a positive tendency with $p <$ .06, and significant correlations for the remaining five accounts range from .206 to .552.

Table 3
Descriptive Statistics for the Liking Experiment

Images	MTurk liking *Mean*	MTurk liking *SD*	IAA scores *Mean*	IAA scores *SD*
Architecture (*n* = 90)	41.24	9.28	−5.27	21.84
Dancer (*n* = 90)	56.80	6.09	−7.24	27.10
Landscape (*n* = 90)	63.28	13.10	−0.52	29.69
Total (*n* = 270)	53.78	13.54	−4.35	26.47

A multiple linear regression to predict Mt\Turk liking by IAA scores and genre revealed a significant relation, $F(2, 267) = 152.04$, $p < .001$, with an adjusted R^2 value of .529. The significant R^2 change for the IAA scores as first predictor is .120. Also, separate multiple regressions for the three genres (90 images each) were calculated, predicting MTurk liking with the account variable and IAA scores. Significant adjusted R squares were found for all genres: architecture, adj. $R^2 = .429$ (R^2 change for IAA scores as first predictor 7.7%); dancer, adj. $R^2 = .330$ (R^2 change for IAA scores as first predictor 33.0%); landscape, adj. $R^2 = .301$ (R^2 change for IAA scores as first predictor 22.7%). See Table 4 for an overview of these R^2 changes and Pearson correlations (MTurk liking and IAA scores) within all nine accounts.

The MTurk liking ratings were subjected to a one-way ANOVA with factor *genre*. It revealed that the mean liking ratings differed between the three genres, $F(2, 267) = 117.45$, $p < .001$. A Tukey post hoc test revealed that MTurk mean liking for architecture ($M = 41.24$, $SD = 9.28$) was significantly lower than that of both dancers ($M = 56.80$, $SD = 6.09$, $p < .001$) and landscapes ($M = 63.28$, $SD = 13.10$, $p < .001$). There was also a significant difference between the ratings for dancers and landscapes, the latter being more appealing ($p < .001$). Looking at IAA scores for the image set used in the experiment, there was no significant difference between genres, $F(2, 267) = 1.539$, $p = .216$.

Discussion

What do these results tell us about the prospects and limitations of IAA scores? First, Pearson correlations between IAA scores and MTurk liking ratings are positive for all nine accounts under investigation. However, they are only low to medium sized, and for two accounts (A2 and D1) they are not significant. Grouping the accounts based on genre, IAA scores explain different propor-

Table 4

Investigating the Relation of MTurk Liking With IAA Scores

Genre	Pearson *r* within accounts	Pearson *r* within genres	Pearson *r* across genres	Predicting MTurk liking: R^2 changes within genres	Predicting MTurk liking: R^2 changes across genres
Architecture	A1 .125 (p = .055) A2 .046 n.s. A3 .129 (p = .052)	.277**		IAA: .077** Account: .365**	
Dancer	D1 .030 n.s. D2 .552** D3 .500**	.575**	.347**(partial *r* when controlling for genre: .401**)	IAA: .330** Account: .005	IAA: .120** Genre: .412**
Landscape	L1 .206* L2 .380** L3 .320**	.477**		IAA: .227** Account: .089**	

Note. Left column: Pearson correlations within accounts. Middle and right column: R^2 changes of multiple linear regression within genres (two predictors: IAA + account) and across genres (two predictors: IAA + genre).
p values for Pearson correlations and R^2 changes reach significance at * $p < .05$ or ** $p < .01$.

tions of variance of the MTurk liking ratings within each of the three genres. In architecture only 8% of the variance can be explained, coupled with a strong effect of the account variable indicating large differences in MTurk liking between the three accounts. The account effect is smaller for the dancer and landscape genres, and this goes hand in hand with much stronger relations with the IAA scores, explaining 33% and 23% of the variance in MTurk liking ratings, respectively.

Before interpreting these numbers, it must be noted that the experimental data only reflect liking of the plain images, whereas Instagram Likes are a result of the image and its complex Instagram context, with captions, hashtags, and potential seasonal or other time-based trends. For reasons of simplicity, we excluded all this information in the experiment and used the isolated images only. Also, the sample size of our random subset of images for the experimental liking task was rather small with only 30 images from each account. To derive more definite statements from the differences between genres and accounts, experiments with larger image samples might be needed. In this first approach, however, we find the effects in dancer and landscape genres sufficient to propose that IAA scores measure similar aspects as controlled experiments asking for liking on a visual analogue scale.

Second, we found general genre preferences in the MTurk liking data, namely the preference for landscapes over dancers and of dancers over architecture. These preferences, however, are not visible in the IAA scores. This is due to a clear methodological limitation: IAA scores result from separate computations per account, and the resulting scores are therefore not useful to compare the scores of different accounts to one another. Genres also only vary across different accounts in our dataset. This explains why genre preferences are not present in the IAA scores. As we will show next, however, it is possible to investigate any image feature that varies within accounts.

Investigating Aesthetic Principles With IAA Scores

To examine the potential of IAA scores, we tested how well they reflect established aesthetic principles. To do so we explored two well-known low-level features (curvature and visual balance), as well as the aesthetic effects of a high-level content variable for one genre (gender of dancers).

Curvature

The preference for curvature as opposed to angularity has been studied systematically over the last decades and the link between curvature and aesthetic preference has been found very consistently across different types of content (Bar & Neta, 2006; Bertamini, Palumbo, Gheorghes, & Galatsidas, 2016; Gómez-Puerto et al., 2016). Most of the studies, however, classify shapes as either curved or angular based on subjective classification without objective quantification. This shortcoming was recently addressed by Grebenkina, Brachmann, Bertamini, Kaduhm, and Redies (2018), who utilized the relation between curvature and the distribution of luminance edge orientations. If a picture is composed of curved elements, then the distribution of edge orientations is usually distributed evenly. From this it is inferred that the more evenly distributed the edge-orientations in a picture are, the more curved the composition appears (which, however, must not necessarily be the case). This so-called *edge-orientation entropy* was found, for instance, to be higher in traditional artworks as compared with several categories of other nonart images (Redies, Brachmann, & Wagemans, 2017).

Method

For measuring curvedness, we computed the first-order edge-orientation entropy for all images in our database, where we applied the same method as Redies, Brachmann, and Wagemans (2017). that is, the grayscale version of each picture was filtered by a set of 24 oriented Gabor filters representing a full circle. The outputs of these filters were then used to compute the Shannon entropy, which was taken as a measure for curvedness of an image. See Redies et al. (2017) for more information on the computation.

Results and Discussion

We found high edge-orientation entropy scores for the majority of our stimulus set of high-quality photographs (see histogram in Figure 6), adding evidence to the finding of Redies et al. (2017) that visually appealing images generally reach quite high entropy scores. Independent-samples *t* tests were conducted to compare mean IAA scores in the lowest and highest edge-orientation entropy quartile within each genre. For all three genres IAA scores were significantly higher for the highest edge-orientation entropy quartile; architecture $M_{\text{low}} = -2.83$ versus $M_{\text{high}} = 0.87$, $t(2049) = -3.19$, $p = .001$, Cohen's $d = 0.141$; dancer $M_{\text{low}} = -7.24$ versus $M_{\text{high}} = -0.07$, $t(1188) = -3.88$, $p < .001$, Cohen's $d = 0.225$; landscape $M_{\text{low}} = -10.49$ versus $M_{\text{high}} = 4.06$, $t(2996) = -15.04$, $p < .001$, Cohen's $d = 0.549$. Figure 6 illustrates these findings and Table 5 sums up the corresponding test statistics and effect sizes. We interpret these findings as evidence that the preference for curvature is reflected by the IAA scores, both within and across the three genres under investigation.

Balance

A frequently used measure of visual balance is the deviation of the center of mass from the geometrical center of the frame (Ross, 1907). It is assumed that the smaller the deviation of the center of mass (DCM) the better a picture is balanced—at least in square format. Indeed, for images with simple geometric forms it has been shown that those with small DCM scores are preferred (Hübner & Fillinger, 2016). For architecture photography, it has been shown that DCM scores could not predict the liking of images representing facade patterns of low complexity. However, for images of higher complexity, DCM scores were successful at predicting liking (Thömmes & Hübner, 2018). For landscape photography, Svobodova, Sklenicka, Molnarova, and Vojar (2014) found that the position of the horizon affects preference. Photos with a horizon located in the middle or upper third of the image were preferred over those with a horizon in the lower third of the image. Svobodova et al. (2014) did not use measures of balance, but the preference they describe does not suggest a preference for well-balanced compositions toward the geometrical center of the image.

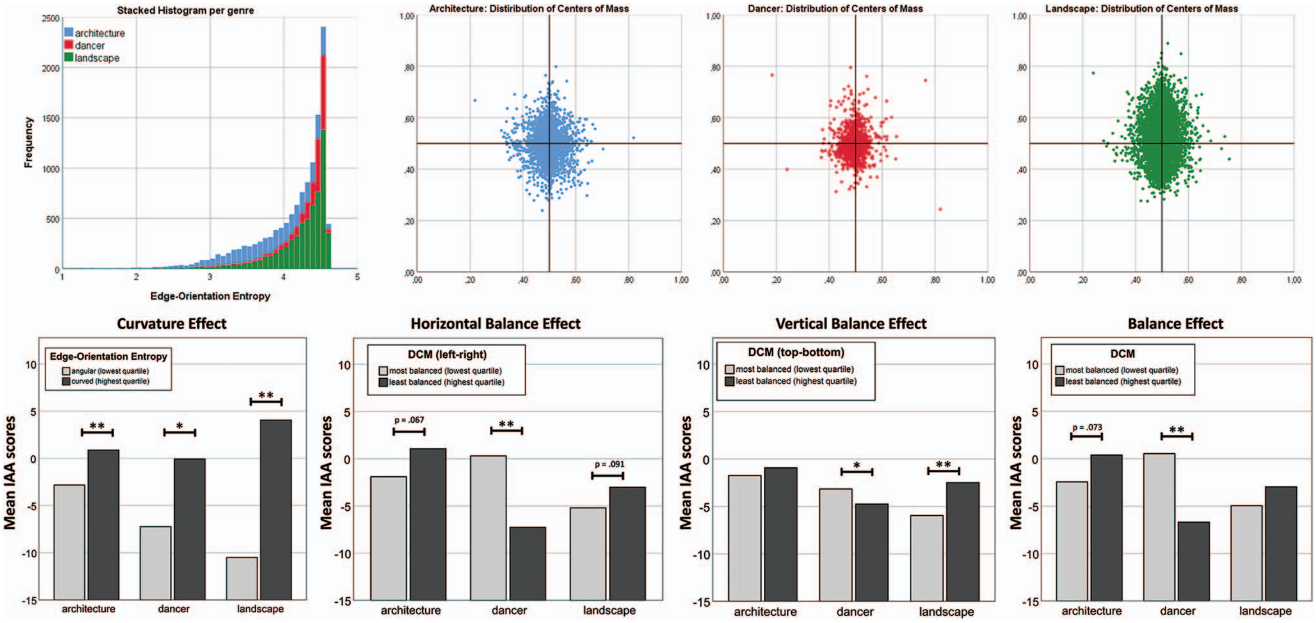


Figure 6. Objective measures of curvature (edge-orientation entropy) and visual balance toward the center of the frame (DCM). Top row from left to right: (1) Stacked Histogram of Edge-Orientation Entropy scores per genre (2–4). Distribution of Centers of Mass per genre, each dot indicating the center of mass in one image (frame standardized to width and height of 1). Bottom row from left to right: (1–4) Mean IAA scores for lowest versus highest quartile of edge-orientation entropy, balance toward the horizontal midline (DCM left-right), balance toward the vertical midline (DCM top-bottom), and balance toward the center of the frame (DCM). Differences reach significance at $^{*}\,p < .05$ or $^{**}\,p < .01$. See the online article for the color version of this figure.

Table 5
Overview of Descriptive Statistics and Quartile Comparisons for Curvature and Balance Measures

	Descriptives			Quartile *Means*		Lowest quartile		Highest quartile			
	N	*M*	*SD*	*lowest*	*highest*	*M*	*SD*	*M*	*SD*	*t*	*Cohen's d*
	Edge-orientation entropy					Compare IAA scores of lowest versus highest OE quartile					
Architecture	4,100	3.800	0.569	3.017	4.437	−2.829	24.317	0.870	28.075	-3.189^{**}	0.141
Dancer	2,377	4.300	0.384	3.794	4.545	−7.235	32.054	−0.071	31.720	-3.875^{**}	0.225
Landscape	5,994	4.241	0.393	3.703	4.554	−10.494	26.794	4.058	26.173	-15.041^{**}	0.549
All	12,471	4.107	0.505	3.372	4.540	−4.988	26.540	2.147	28.671	-10.201^{**}	0.258
	DCM (square format only)					Compare IAA scores of lowest versus highest DCM quartile					
Architecture	3,721	7.530	6.397	1.393	16.70	−2.291	25.335	−0.080	27.806	−1.797	n.s.
Dancer	1,407	7.077	5.714	2.213	14.48	0.151	31.302	−10.409	32.809	4.378^{**}	0.329
Landscape	4,639	12.245	7.518	4.075	22.61	−4.642	27.609	−2.876	29.687	−1.484	n.s.
All	9,767	9.704	7.280	2.304	19.94	−2.769	27.317	−2.320	29.646	−0.550	n.s.
	Horizontal DCM (square format only)					Compare IAA scores of lowest versus highest DCM_{hori} quartile					
Architecture	3,721	4.779	6.018	0.351	13.31	−1.941	26.250	0.373	28.217	−1.821	n.s.
Dancer	1,407	4.770	5.342	0.754	11.31	−1.523	33.706	−9.645	31.235	3.343^{**}	0.250
Landscape	4,639	6.033	5.692	0.953	13.80	−4.794	27.222	−2.826	28.149	−1.692	n.s.
All	9,767	5.373	5.804	0.636	13.21	−2.681	27.484	−2.313	29.026	−0.460	n.s.
	Vertical DCM (square format)					Compare IAA scores of lowest versus highest $\text{DCM}_{\text{verti}}$ quartile					
Architecture	3,721	8.139	8.364	0.839	20.01	−1.781	24.595	−1.268	27.414	−0.424	n.s.
Dancer	1,407	7.654	7.456	1.154	17.83	−2.863	28.589	−8.333	33.418	2.288^{*}	0.176
Landscape	4,639	14.867	11.116	3.041	30.35	−5.589	28.058	−2.107	29.729	-2.893^{**}	0.155
All	9,767	11.264	10.242	1.635	26.08	−3.028	27.290	−1.765	29.509	−1.570	n.s.

Note. On the left there are descriptive statistics of all low-level features under investigation, separately for all genres. On the right the corresponding test statistics and effect sizes can be found. *p*-values for *t*-tests reach significance at $^{*}\,p < .05$ or $^{**}\,p < .01$.

As far as we know, there has been no research on visual balance in dancer photography, yet. The dancer photographs in our database are staged compositions, focusing on the dancers' bodies and their movements before a neutral background. As these photographs are thus relatively standardized, we expected differences in visual balance to play a pronounced role for their aesthetic appeal.

Method

The DCM scores are computed based on pixel luminance (Hübner & Fillinger, 2016; Thömmes & Hübner, 2018). For the current investigation, we computed three measures representing the horizontal, vertical, and absolute deviation of the center of mass from the geometrical midline or center, respectively: DCM_{hori} (percentage left-right deviation from vertical midline), $\text{DCM}_{\text{verti}}$ (percentage top-down deviation from horizontal midline), DCM (percentage deviation from geometrical midpoint). For further details of the formulas see Thömmes and Hübner (2018). Importantly, we only computed these measures for images in square format, as the particular importance of the geometrical center becomes weaker in nonsquare compositions (Arnheim, 1982) and previous research also focused on the square format with respect to DCM scores. For the architecture genre, there were 3,721 images in square format, and 1,407 and 4,639 images for dancers and landscapes, respectively. We computed the measures differently for high-key and low-key images. That is, if the foreground was dark and the background bright, we inverted the gray levels. Because for computing the DCM it is assumed that each pixel's weight corresponds to its gray level, this inversion ensured that the center of mass was mainly determined by the objects in the foreground (e.g., the dancer) rather than by the background.

Results and Discussion

Table 5 sums up the descriptive statistics for the three visual balance measures. As Figure 6 illustrates, the distribution of centers of mass differs between genres, with dancer portraits being the most balanced toward the geometrical center, landscapes scatter widely along the vertical axis, and the architecture photographs being in between of these two. This suggests that balancing the composition toward the center plays the biggest role in dancer photography.

But how do the DCM scores relate to IAA scores as a measure of the aesthetic appeal of the photographs? Again, independent-samples *t* tests were conducted to compare mean IAA scores in the lowest quartiles and the highest quartiles of the three visual balance measures (DCM, $\text{DCM}_{\text{verti}}$, DCM_{hori}) within each genre (see Table 5). In the architecture genre as a whole, no significant differences were found. There is, however, a tendency for less balanced images to be preferred for DCM_{hori} ($p = .069$) and DCM ($p = .073$), see the corresponding graphs in Figure 6. Based on previous findings on architecture photos, we classified all architecture images as "pattern-like" (low complexity compositions, rotation does not make the image appear upside down) or "scene-like" compositions (higher complexity, clearly perceivable top and bottom) based on rotation invariance of the image (cf. Thömmes & Hübner, 2018). For "pattern-like" images we found the preference for less balanced images ($M_{low} = -3.14$ vs. $M_{high} = 2.75$, $t(651) = -2.65$, $p = .008$, Cohen's $d = 0.229$); however, in the "scene-like" images balance had no significant effect on IAA scores. The left graph in Figure 7 illustrates this. This replicates the findings of Thömmes and Hübner (2018) only in part. For "pattern-like" image compositions the preference for less balanced images was replicated. However, for more complex "scene-like" compositions the positive effect of visual balance could not be generalized to the present pictures. It might be the case that the relation between balance and liking is nonlinear in the sense that small deviations are preferred over perfect centering. An inverted u-shape relation between the DCM and aesthetic appeal was already discussed in Thömmes and Hübner (2018). Moreover, such a relation has also been suggested for the interplay between complexity and liking (Imamoglu, 2000).

In the dancer genre, there were significant differences in IAA scores in all three measures with higher IAA scores for the lowest DCM quartile (more balanced); DCM $M_{low} = 0.15$ versus $M_{high} = -10.41$, $t(705) = 4.38$, $p < .001$, Cohen's $d = 0.329$; DCM$_{\mathrm{hori}}$ $M_{low} = -1.52$ versus $M_{\mathrm{high}} = -9.65$, $t(714) = 3.34$, $p = .001$, Cohen's $d = 0.250$; DCM$_{\mathrm{verti}}$ $M_{low} = -2.86$ versus $M_{\mathrm{high}} = -8.33$, $t(676) = 2.29$, $p = .022$, Cohen's $d = 0.176$. The situation here is very conclusive: For square photos in the dancer genre, balanced pictures were clearly preferred.

For landscapes, significant effects on IAA scores were only found for the DCM$_{\mathrm{verti}}$ measure. The independent-samples t test comparing the mean IAA scores in the lowest quartile ($M_{low} = -5.59$) with that in the highest quartile of DCM$_{\mathrm{verti}}$ ($M_{high} = -2.11$) revealed that less balanced images were significantly more appealing, $t(2306) = -2.89$, $p = .004$, Cohen's $d = 0.155$. This implies a preference for centers of mass that deviate from the horizontal midline, which might explain the additional variance in the centers of mass along the vertical midline in the landscape genre compared to other genres (see graph in the upper right in Figure 6). This result might also give point to Svobodova et al.'s (2014) findings regarding the locations of the horizon, however, a preference for deviation from the center does not allow definite statements on preferences for specific locations regarding the rule of thirds or golden ratios.

Gender

So far, we found that both considered low-level features of composition affected the aesthetic appeal of the images in the dancer genre. To evaluate the relative size of these effects, we also investigated the effect of the high-level content variable gender of the dancing person. We hypothesized that female dancers are preferred over male dancers in all three dancer accounts, as females are not only the prototype of a ballet dancer, but also wear more colorful and flamboyant dresses in many photos compared with males.

Method

We classified all 2,378 images from the dancer genre based on the gender of the dancer(s) resulting in four groups: female(s) ($n = 1626$), male(s) ($n = 417$), mixed gender couples or groups ($n = 253$), and other content ($n = 82$). We excluded 82 images that did not contain dancers. For the remaining images, we investigated gender effects on IAA scores.

Results and Discussion

IAA scores were subjected to a one-way ANOVA with the factor gender. It revealed that mean IAA scores differed between

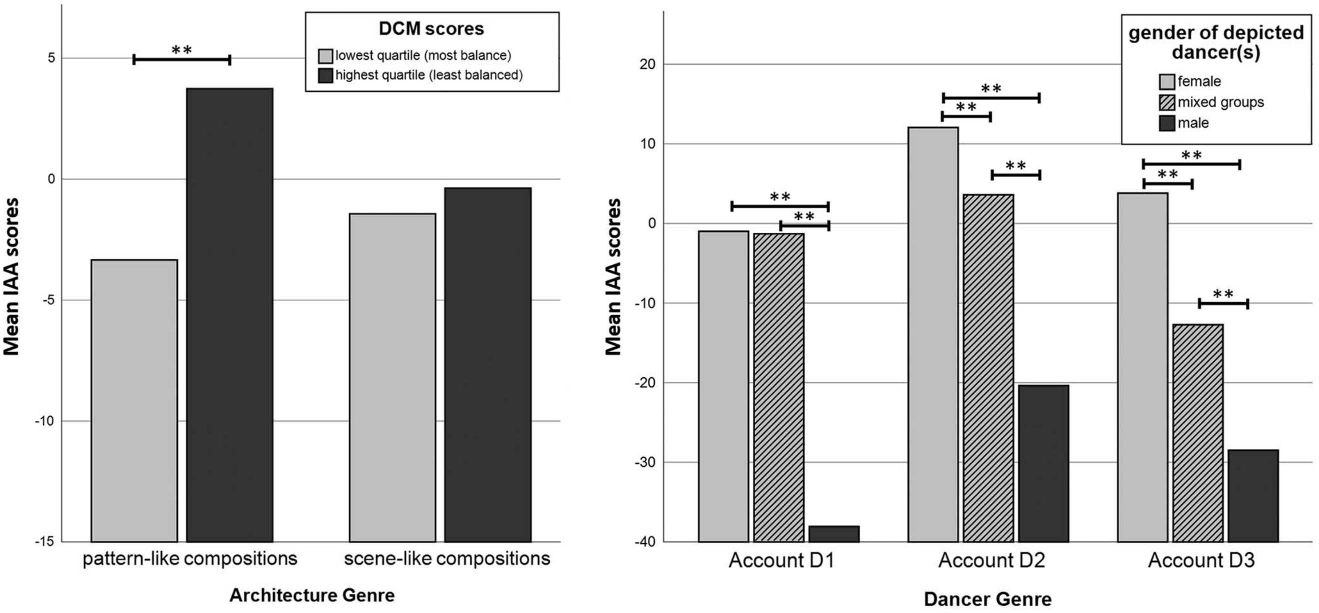


Figure 7. The left graph illustrates the balance effect on IAA scores in the architecture genre after splitting the genre into images with pattern-like versus scene-like composition. The right graph illustrates the gender effect on IAA scores that is present in all three dancer accounts. Differences reach significance at $^{*}p < .05$ or $^{**}p < .01$.

the three categories, $F(2, 2293) = 152.17$, $p < .001$. A Tukey post hoc test revealed that IAA scores for female dancers ($M = 3.81$, $SD = 31.26$) are significantly higher than both mixed gender groups ($M = -4.71$, $SD = 33.65$, $p < .001$, Cohen's $d = 0.262$), and male dancers ($M = -24.96$, $SD = 22.41$, $p < .001$, Cohen's $d = 1.058$). Also, mixed gender groups were aesthetically more appealing than male dancer(s), $p < .001$, Cohen's $d = 0.708$. The right graph in Figure 7 illustrates these gender effects, being present in all three dancer accounts when analyzing them separately, with only one nonsignificant difference between female(s) and mixed groups for account D1. This provides strong evidence that the gender of the dancer affects the aesthetic appeal of dancer portraits, with lowest IAA scores for male dancers, higher IAA scores as soon as at least one woman is present and even higher IAA scores when no male dancer is present.

Discussing the Empirical Validation

Taken together, we found evidence that the preference for curvature (as measured with the edge-orientation entropy) is an aesthetic principle present in all the image types in our database (curvature Cohen's d ranging from 0.141 to 0.549). For visual balance the relationships were more complex and domain-specific, and we suppose that higher standardization of the image set is necessary in order to reveal balance effects. We found a preference for well-balanced compositions in square dancer photographs. However, for both architecture and landscape photographs our findings suggest that less balance is preferred, which might give point to other photographic rules (such as the golden ratio, see Svobodova et al., 2014) when composing pleasing image compositions in those genres (visual balance Cohen's d ranging from 0.155 to 0.329). The gender of the dancer as a more obvious content variable was affecting IAA scores more pronouncedly, female dancers being aesthetically more appealing than male dancers (male vs. female dancers Cohen's $d = 1.058$).

Another interesting approach that becomes possible using the IAA measure as a proxy for aesthetic appeal, is to *compare effect sizes between different genres*. Comparing the curvature effect for all three genres reveals some differences. Curvature affects IAA scores most strongly in the landscape genre (Cohen's $d = 0.549$) and less strongly in dancers (Cohen's $d = 0.225$) and architecture (Cohen's $d = 0.141$). This might be a starting point for deeper investigation of the fundamental causes of such differences. For the dancer genre it is also interesting to *compare the three effects within genre*, with gender being the largest (Cohen's $d = 1.058$), visual balance (Cohen's $d = 0.329$) and curvature (Cohen's $d = 0.225$) being smaller in comparison. To bring it all together, we conducted a multiple linear regression comparing the relative importance of these three factors in the dancer genre. After excluding nondancer content ($n = 82$) and nonsquare format images ($n = 971$), for the remaining 1,345 images a multiple linear regression was calculated to predict IAA scores based on the gender of the dancer (male = 1, mixed = 2, female = 3), balance (DCM scores), and curvature (edge-orientation entropy). A significant regression equation was found, $F(3, 1340) = 99.38$, $p < .001$, with an adjusted R^2 of .180. Gender as first predictor accounts for 13.4% of the variance, balance as second predictor for an additional 3.3%, and curvature as third predictor additional 1.4% (all R^2 changes reach significance at $p < .01$). These numbers give us an idea of the importance of the three factors relative to one another. It is obvious that the content factor gender has the largest effect on aesthetic appeal, while balance and curvature play a much smaller role. However, balance is more than twice as important as curvature, which makes sense in view of the types of dancer images. When it comes to evaluating the relative importance of aesthetic effects (especially when low-level features are concerned), Matz, Gladstone, and Stillwell (2017) raise the vital question how one distinguishes between small effects that are meaningful and those that are not. For broad and large field studies such as ours, they stress the importance of even small effects in the data (Matz et al., 2017),[5] especially when hypothesizing happens with previous research findings in mind. IAA scores then are a great way to generalize observations from the lab to real-life stimuli.

General Discussion

The aim of the present study was to introduce a measure that can be used as a proxy for the aesthetic appeal of photographs. Based on Instagram Likes for photos of professional photographers we developed and validated the IAA score. This measure effectively discounts confounding effects of time and growing numbers of followers. We expect the introduced method to work for most large accounts (>10,000 followers) that are monothematic with a frequent posting habit. To date, we have only investigated professional photography from three genres (architecture, dancer, landscape). How well the introduced method works for other content on Instagram remains to be investigated in the future.

There is evidence that IAA scores reliably measure the aesthetic appeal of images on Instagram. Due to the uncontrolled nature of the data, it might even seem surprising that we found good test–retest reliability for images posted twice on the same account. However, we have this data only for one account (D2) which is also one of the most consistent in terms of content. For other accounts with more diverse content, further investigation is needed.

Overall, both experimental and empirical validation confirm IAA scores to be a useful measure of aesthetic appeal. Experimental liking and IAA scores are closely related in the landscape and dancer genre. For architecture photography, however, we found a weaker connection between experimental liking data and IAA scores. This might reflect the limits of our experimental setup, such as small sample size (30 images per account) and only asking for liking. Using a more comprehensive set of aesthetic descriptors, such as interestingness, beauty, harmony, and pleasingness might result in a better understanding of IAA scores. Such an exploration would be a promising starting point for future research. Nevertheless, we think that the present data are sufficient for deeper analyses of aesthetic features and their effects on the IAA measure. Before investigating aesthetic features in our database, we also discussed another important methodological limitation: IAA scores are computed for each account separately, and there-

[5] "Psychologists have typically focused on how their findings apply to individuals. However, by providing the opportunity to understand and influence the behaviors of billions of people around the world, the era of big data encourages—and possibly requires—researchers to think bigger. In this new world, small effects can still matter." (Matz et al., 2017, p. 550)

fore they are not useful to compare different Instagram accounts to one another. With our database, in which genres only vary between accounts, it is therefore not possible to investigate general genre preferences. Yet, IAA scores are useful to compare any feature that varies within accounts. This was demonstrated by successfully investigating the aesthetic effects of low-level features (curvature, visual balance) as well as of a high-level content variable (gender of dancer). These examples illustrate the potential of our proposed measure of aesthetic appeal.

Taken together, our attempts to empirically validate IAA scores underline two things: First, well-known aesthetic preferences are reflected in the IAA measure, giving point to using it as a proxy for aesthetic appeal. Second, with IAA scores available for large data sets, it becomes possible to compare effects of different features both within and across photographic genres. This prepares ground to investigate different aesthetic features—be it compositional image features, semantic content, or even contextual aspects such as hashtags and caption texts on Instagram—relative to one another and to get closer to solving the aesthetic riddle with complex real-world stimuli. We want to encourage further research to explore the scope of IAA scores as measure of aesthetic appeal and to demonstrate its prospects and limitations.

References

Abutaleb, Y. (2016, June 21). *Instagram's user base grows to more than 500 million* [Blog post]. Retrieved from https://www.reuters.com/article/us-facebook-instagram-users/instagrams-user-base-grows-to-more-than-500-million-idUSKCN0Z71LN

Anna. (2018, April 4) *Updates to Instagram data and Minter.io* [Blog post]. Retrieved from https://help.minter.io/platform-updates/updates-to-instagram-data-and-minterio

Arnheim, R. (1982). *The power of the center: A study of composition in the visual arts*. Berkeley, CA: University of California Press.

Augustin, M. D., Wagemans, J., & Carbon, C.-C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica, 139,* 187–201. http://dx.doi.org/10.1016/j.actpsy.2011.10.004

Axelsson, O. (2007). Towards a psychology of photography: Dimensions underlying aesthetic appeal of photographs. *Perceptual and Motor Skills, 105,* 411–434. http://dx.doi.org/10.2466/pms.105.2.411-434

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science, 17,* 645–648. http://dx.doi.org/10.1111/j.1467-9280.2006.01759.x

Berlyne, D. E. (1971). *Aesthetics and psychobiology. The century psychology series*. New York, NY: Meredith.

Bertamini, M., Palumbo, L., Gheorghes, T. N., & Galatsidas, M. (2016). Do observers like curvature or do they dislike angularity? *British Journal of Psychology, 107,* 154–178. http://dx.doi.org/10.1111/bjop.12132

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the Effect Size of d for Range Restriction and Unreliability. *Organizational Research Methods, 4,* 46–61. http://dx.doi.org/10.1177/109442810141003

Carroll, H. (2017). *Read this if you want to be instagram famous: 50 Secrets by 50 of the best*. London, UK: Laurence King Publishing.

Cerosaletti, C. D., & Loui, A. C. (2009). Measuring the perceived aesthetic quality of photographic images. In *QoMEx 2009 International Workshop on Quality of Multimedia Experience* (pp. 47–52). San Diego, CA: IEEE. http://dx.doi.org/10.1109/QOMEX.2009.5246977

Denzler, J., Rodner, E., & Simon, M. (2016). Convolutional neural networks as a computational model for the underlying processes of aesthetics perception. In G. Hua & H. Jégou (Eds.), *European Conference on Computer Vision* (pp. 871–887). Cham, Switzerland: Springer.

Fechner, G. T. (1876). *Vorschule der Aesthetik* [Groundwork on Aesthetics]. Weimar, Leipzig: Universitätsbibliothek; Breitkopf & Härtel.

Ferrara, E., Interdonato, R., & Tagarelli, A. (2014). Online popularity and topical interests through the lens of instagram. In L. Ferres, G. Rossi, V. Almeida, & E. Herder (Eds.), *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (pp. 24–34). New York, NY: Association for Computing Machinery (ACM).

Gan, C. (2017). Understanding WeChat users' liking behavior: An empirical study in China. *Computers in Human Behavior, 68,* 30–39. http://dx.doi.org/10.1016/j.chb.2016.11.002

Gershoni, S., & Kobayashi, H. (2006). How we look at photographs as indicated by contrast discrimination performance versus contrast preference. *Journal of Imaging Science and Technology, 50,* 320–326. http://dx.doi.org/10.2352/J.ImagingSci.Technol.(2006)50:4(320)

Gómez-Puerto, G., Munar, E., & Nadal, M. (2016). Preference for curvature: A historical and conceptual framework. *Frontiers in Human Neuroscience, 9,* 712. http://dx.doi.org/10.3389/fnhum.2015.00712

Grebenkina, M., Brachmann, A., Bertamini, M., Kaduhm, A., & Redies, C. (2018). Edge-orientation entropy predicts preference for diverse types of man-made images. *Frontiers in Neuroscience, 12,* 678. http://dx.doi.org/10.3389/fnins.2018.00678

Hübner, R., & Fillinger, M. G. (2016). Comparison of objective measures for predicting perceptual balance and visual aesthetic preference. *Frontiers in Psychology, 7,* 335. http://dx.doi.org/10.3389/fpsyg.2016.00335

Imamoglu, C. (2000). Complexity, liking and familiarity: Architecture and non-architecturs Turkish students' assessment of traditional and modern house facades. *Journal of Environmental Psychology, 20,* 5–16. http://dx.doi.org/10.1006/jevp.1999.0155

Instagram Press. (2018, March 22). *Changes to improve your Instagram feed* [Blog post]. Retrieved from https://about.instagram.com/blog/announcements/new-posts-button-to-improve-instagram-feed

Jacobsen, T., Buchta, K., Köhler, M., & Schröger, E. (2004). The primacy of beauty in judging the aesthetics of objects. *Psychological Reports, 94*(3 Part 2), 1253–1260. http://dx.doi.org/10.2466/pr0.94.3c.1253-1260

Jang, J. Y., Han, K., & Lee, D. (2015). No reciprocity in "liking" photos: Analyzing like activities in Instagram. In Y. Yesilada, R. Farzan, & G.-J. Houben (Chairs), *The 26th ACM Conference,* Guzelyurt, Northern Cyprus.

Komok, A. (2018a, June 19) *What is Instagram engagement rate and how to calculate it* [Blog post]. Retrieved from https://hypeauditor.com/blog/what-is-instagram-engagement-rate-and-how-to-calculate-it/

Komok, A. (2018b, March 28) *How do micro-influencers and mega-influencers compare in Instagram engagement rates* [Blog post]. Retrieved from https://medium.com/influencer-marketing-made-easy/how-do-micro-influencers-and-mega-influencers-compare-in-instagram-engagement-rates-cfab691ed600

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision ECCV 2016 Proceedings Part I* (pp. 662–679). Cham, Switzerland: Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15,* 155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Lebreton, P., Raake, A., & Barkowsky, M. (2016). Evaluation of aesthetic appeal with regard of user's knowledge. Electronic Imaging: Human Vision and Electronic Imaging 2016, Springfield, VA, (16), 1–6. http://dx.doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-119

Lee, D. (2016). Likeology: Modeling, predicting, and aggregating likes in social media. In W. Neijdl & W. Hall (Eds.), *Proceedings of the 8th ACM Conference on Web Science* (p. 13). New York, NY: Association for Computing Machinery (ACM).

Lowe-Calverley, E., & Grieve, R. (2018). Thumbs up: A thematic analysis of image-based posting and liking behaviour on social media. *Telematics and Informatics, 35,* 1900–1913. http://dx.doi.org/10.1016/j.tele.2018.06.003

Matz, S. C., Gladstone, J. J., & Stillwell, D. (2017). In a world of big data, small effects can still matter: A Reply to Boyce, Daly, Hounkpatin, and Wood (2017). *Psychological Science, 28,* 547–550. http://dx.doi.org/10.1177/0956797617697445

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46. http://dx.doi.org/10.1037/1082-989X.1.1.30

McManus, I. C., Stöver, K., & Kim, D. (2011). Arnheim's Gestalt theory of visual balance: Examining the compositional structure of art photographs and abstract images. *I-Perception, 2,* 615–647. http://dx.doi.org/10.1068/i0445aap

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2408–2415). Washington, DC: IEEE. Retrieved from https://dl.acm.org/doi/proceedings/10.5555/2354409

Redies, C., Brachmann, A., & Wagemans, J. (2017). High entropy of edge orientations characterizes visual artworks from diverse cultural backgrounds. *Vision Research, 133,* 130–144. http://dx.doi.org/10.1016/j.visres.2017.02.004

Ross, D. W. (1907). *A theory of pure design: Harmony, balance and rhythm*. Boston, MA: Houghton, Mifflin and Company.

Russell, P. A., & George, D. A. (1990). Relationships between Aesthetic Response Scales Applied to Paintings. *Empirical Studies of the Arts, 8,* 15–30. http://dx.doi.org/10.2190/AU1R-6UXE-T14R-04WQ

Siahaan, E., Hanjalic, A., & Redi, J. (2016). A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Transactions on Multimedia, 18,* 1338–1350. http://dx.doi.org/10.1109/TMM.2016.2559942

Svobodova, K., Sklenicka, P., Molnarova, K., & Vojar, J. (2014). Does the composition of landscape photographs affect visual preferences? The rule of the Golden Section and the position of the horizon. *Journal of Environmental Psychology, 38,* 143–152. http://dx.doi.org/10.1016/j.jenvp.2014.01.005

Systrom, K. (2018, June 20). *Welcome to IGTV* [Blog post]. Retrieved from https://about.instagram.com/blog/announcements/welcome-to-igtv

Thömmes, K., & Hübner, R. (2018). Instagram likes for architectural photos can be predicted by quantitative balance measures and curvature. *Frontiers in Psychology, 9,* 1050. http://dx.doi.org/10.3389/fpsyg.2018.01050

Tinio, P. P. L., Leder, H., & Strasser, M. (2011). Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together. *Psychology of Aesthetics, Creativity, and the Arts, 5,* 165–176. http://dx.doi.org/10.1037/a0019542

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. Cognition, 179, 121–131. http://dx.doi.org/10.1016/j.cognition.2018.06.009

Vessel, E. A., Stahl, J., Maurer, N., Denker, A., & Starr, G. G. (2014). Personalized visual aesthetics. In B. E. Rogowitz, T. N. Pappas, & H. de Ridder (Eds.), *Proceedings. SPIE 9014, Human Vision and Electronic Imaging XIX, 90140S* (pp. 1–8). San Francisco, CA: SPIE. http://dx.doi.org/10.1117/12.2043126

Wilson, A., & Chatterjee, A. (2005). The assessment of preference for balance: Introducing a new test. *Empirical Studies of the Arts, 23,* 165–180. http://dx.doi.org/10.2190/B1LR-MVF3-F36X-XR64

Appendix A

Descriptive Statistics of the Instagram Database

We gathered the following data for every image in our database: number of Likes, number of comments, time created, historical followers (number of followers when image was posted), and caption text (possibly also containing hashtags and links). We also collected descriptive account data for the nine accounts. This table sums up descriptive statistics.

							Likes[a]		Comments[a]	
	Content	Instagram name	# images[a]	# followers[a]	First post	Account age [days since first post][a]	*Mean*	*SD*	*Mean*	*SD*
A1	architecture	@fernsehturm_	1,321	18,249	11/2014	1208	866.84	390.19	32.77	21.63
A2	architecture	@le_blanc	695	203,239	12/2010	2628	2632.01	1493.10	76.56	40.36
A3	architecture	@Macenzo	2,329	424,335	06/2011	2486	4708.45	2800.66	86.99	47.23
D1	dancer	@karolinakuras	1,147	42,029	01/2012	2252	1377.69	1023.24	11.65	10.26
D2	dancer	@NYCdanceproject	844	186,256	01/2014	1534	4530.30	2239.72	25.88	23.07
D3	dancer	@rachelnevillephoto	980	94,295	03/2014	1451	2745.97	2011.02	23.52	20.68
L1	landscape	@janske	362	241,557	03/2011	2554	3398.40	2077.52	181.16	86.82
L2	landscape	@jn	6,595	547,475	10/2010	2705	6149.28	3984.50	73.13	57.01
L3	landscape	@othellonine	800	182,044	07/2011	2427	2632.73	2164.23	53.25	33.81
total			15,073							

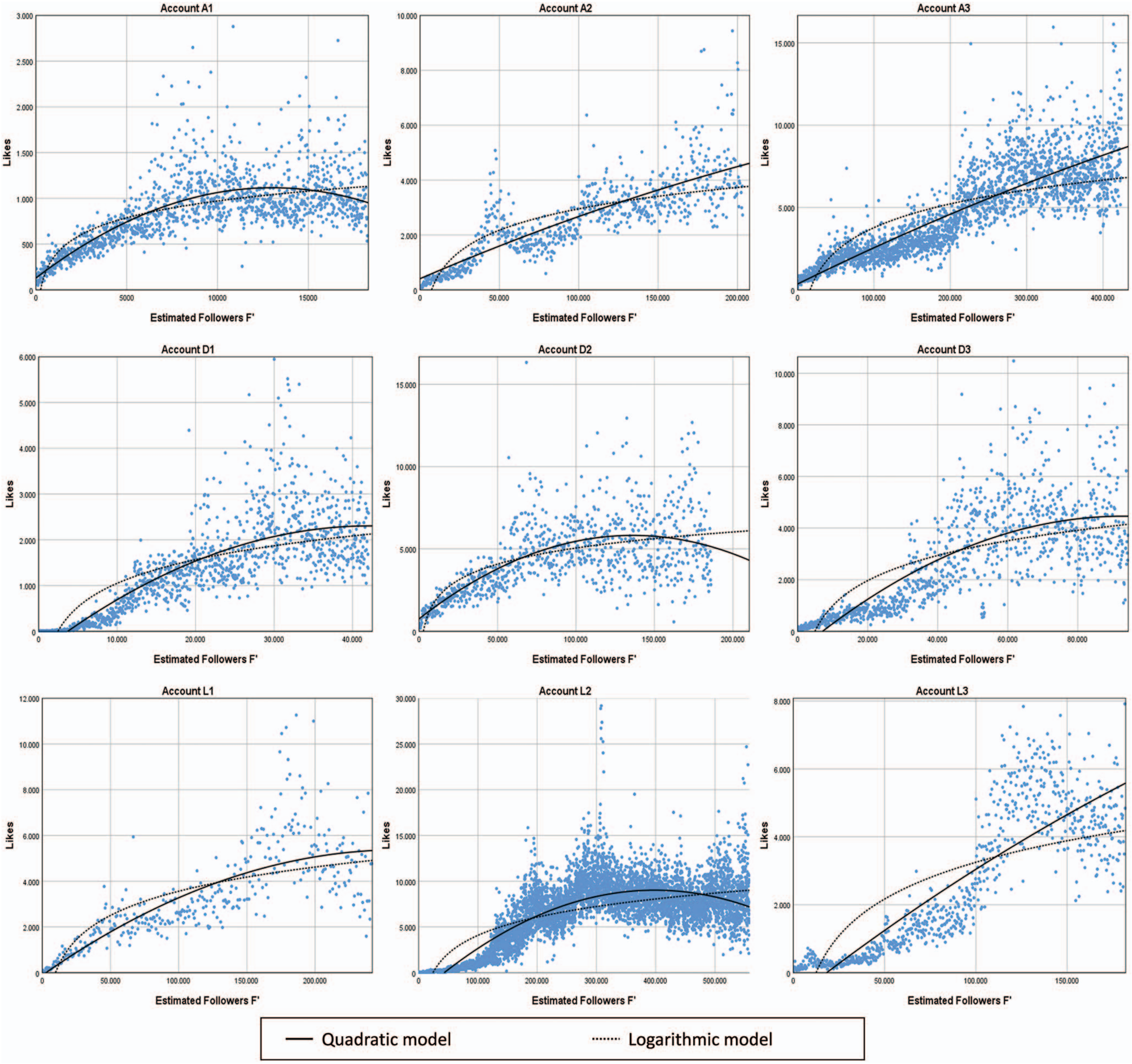
[a] Data collected in April 2018.

(*Appendices continue*)

Appendix B

Numbers of Likes as a Function of Estimated Followers (F')

Two regression models are tested: quadratic and logarithmic prediction of Likes by estimated followers (F'). The following graphs illustrate the data for all nine accounts separately.



Received October 29, 2019
Revision received February 25, 2020
Accepted April 14, 2020 ■