Empirical Studies of the Arts
0(0) 1–19
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0276237418805656
journals.sagepub.com/home/art
SAGE

Relations Between Balance, Prototypicality, and Aesthetic Appreciation for Japanese Calligraphy

Martin G. Fillinger[1] and
Ronald Hübner[1]

Abstract

Aesthetic appreciation of pictures partly depends on the perceptual balance of their elements. This relation has also been supported by objective measures predicting balance ratings as well as preference. Gershoni and Hochstein, however, applied these measures to Japanese calligraphies and failed to find such a relation, which questions the generality of these balance concepts. In our first experiment, we, therefore, tried to replicate these results with a slightly different method. In addition, we calculated further balance measures and collected liking ratings. As result, perceptual balance was again uncorrelated with the measures and with liking. In a second experiment, participants assessed the perceptual stability of the calligraphies, which was considered as alternative concept of balance, and their prototypicality. After discounting the effects of prototypicality on liking, there were correlations between liking and stability and between liking and one of the balance measures. However, the correlations were reliable only for atypical calligraphies.

Keywords

empirical aesthetics, preference, prototypicality, visual stability, perceptual balance

[1] Universität Konstanz, Germany

Corresponding Author:
Martin G. Fillinger, Universität Konstanz, Fachbereich Psychologie, Fach D29. D-78457 Konstanz, Germany.
Email: martin.fillinger@uni-konstanz.de

Introduction

The aesthetic appreciation of pictures depends on perceptual factors such as form, color, and symmetry as well as on more conceptual factors like semantic content and prototypicality (for an overview, see Palmer, Schloss, & Sammartino, 2013). A perceptual factor that has received great attention in art theory and related fields is pictorial balance, that is, how well the picture's elements are arranged (Arnheim, 1982; Bouleau, 1980; Kandinsky, 1926/1979). A picture is commonly considered as balanced if the perceptual "weights" of the pictorial elements are in equilibrium. In recent years, computational measures have even been proposed for representing the balance of a picture. One is the *Assessment of Preference for Balance* (APB), suggested by Wilson and Chatterjee (2005). The APB score is defined as the average of eight symmetry measures over the four axes of a picture (horizontal, vertical, and the two diagonals). As an alternative measure of balance, Hübner and Fillinger (2016) proposed the *Deviation of the Center of "Mass"* (DCM) from the picture's geometrical center. The DCM is based on Arnheim's (1954) idea that the more the center of perceptual "mass" deviates from the geometric center of the picture, the less the picture is perceived as balanced.

Results obtained with the APB and DCM show that these measures do not only correlate with balance ratings but also predict liking ratings. However, the pictures applied in these studies were relatively simple and included only basic (e.g. circles, or squares) and unrelated geometrical elements, which raises the question to what extent the results can be generalized. In an attempt to answer this question, McManus, Stöver, and Kim (2011) could show that the center of mass in art photographs deviated less from the image center than that in random photographs. Yet, there is also negative evidence. Gershoni and Hochstein (2011), for instance, applied the APB to Japanese calligraphies and found no substantial correlation with balance ratings. These results suggest that the proposed formal balance measures are not valid for all picture types. Moreover, because the participants were obviously able to rate perceptual balance, it seems that there are concepts of balance not reflected by the formal measures. If this were indeed the case, then it would be important to know the different concepts of balance and the respective relevant stimulus features.

The aim of the present study was to investigate these issues. In a first step, we wanted to replicate and extend the results of Gershoni and Hochstein (2011). These researchers presented their stimuli very briefly (200 ms) and limited processing duration by masking. Moreover, the stimuli occurred at a random position (spatial uncertainty). Although it is known that balance can be perceived rapidly (Locher, Krupinski, Mello-Thoms, & Nodine, 2007; Locher & Nagy, 1996; Locher & Stappers, 2002; McManus, Edmondson, & Rodger, 1985), it cannot be excluded that Gershoni and Hochstein's procedure was suboptimal for obtaining balance ratings related to the measures. Therefore, we used the same Japanese calligraphies as in the original study but presented the stimuli for a longer time and without spatial uncertainty.

Moreover, in addition to the APB, we also computed the DCM and homogeneity (HG) scores as measures of balance, where HG reflects the distribution of mass within a picture (Hübner & Fillinger, 2016). Because visual complexity might be related to balance as well (Gartus & Leder, 2017; Jacobsen, 2004), we additionally used the file size (FS) of the jpeg compressed image (Donderi, 2006; Forsythe, Nadal, Sheehy, Cela-Conde, & Sawey, 2011; Machado et al., 2015) and the degree of mirror symmetry (MS) in a picture (Hübner & Fillinger, 2016) as corresponding measures.

Furthermore, we wanted to examine the relation between balance and liking for the Japanese calligraphies. As mentioned in art theory (Arnheim, 1982; Bouleau, 1980; Kandinsky, 1926/1979), pictorial balance is considered as a crucial variable for aesthetic appreciation. This relation has been confirmed by Wilson and Chatterjee (2005) and by Hübner and Fillinger (2016). Surprisingly, Gershoni and Hochstein merely asked participants to assess the balance of the calligraphies, but not how much they liked them. Perhaps they supposed that perceptual balance and liking are generally closely related. To test whether this is indeed the case, we additionally asked our participants to indicate how much they liked the Japanese calligraphies.

As we will show, the balance ratings in our first experiment are similar to those of Gershoni and Hochstein (2011). Accordingly, there were again no substantial relations between the balance ratings and the different measures. This suggests that the concept of balance for Japanese calligraphies differs from that reflected by the measures. To investigate these issues further, we conducted a second experiment in which we examined perceived “stability” (Liu, Dong, Zhang, & Jiang, 2017) as an alternative concept of balance.

A further result of our first experiment was that there was no correlation between balance ratings and liking ratings, which indicates that other variables than balance determined the aesthetic differences between the Japanese calligraphies. To examine a possible variable in this respect, in our second experiment we also tested to what extent prototypicality (Rosch, 1975) determines how much Japanese calligraphies are liked. At least in product design, it has been shown that prototypicality can affect preference judgments (e.g., Hekkert, Snelders, & Van Wieringen, 2003; Whitfield & Slatter, 1979).

Experiment 1

The main goal of our first experiment was to investigate whether the absent relation between balance ratings and APB scores in the study of Gershoni and Hochstein (2011) was due to the specific picture type (Japanese calligraphies) or to the relatively short and spatially uncertain stimulus presentation. Therefore, we presented the stimuli for a longer duration and without spatial uncertainty.

Moreover, we also collected liking ratings for the stimuli to examine whether the widely assumed positive relation between perceptual balance and aesthetic appreciation also holds for Japanese calligraphies. As computational measures for balance, we applied not only APB scores but also DCM and HG scores (Hübner & Fillinger, 2016). Furthermore, for additionally assessing effects of visual complexity, we considered the FS (file size) of our images (Donderi, 2006) and MS (mirror symmetry) scores (Gartus & Leder, 2017; Hübner & Fillinger, 2016) as corresponding measures.

Method

Participants. One hundred forty-nine participants (32 men, mean age 23.6 years, $SD = 6.61$) were recruited via an online system (ORSEE; Greiner, 2015) for participation in an online experiment. The pool of persons includes students of all disciplines from the Universität Konstanz as well as persons not associated with our university. As incentive, participants had the chance to win a voucher for their participation. This study was carried out in accordance with the ethical guidelines of the Universität Konstanz and the Declaration of Helsinki. Participants were informed of their right to abstain from participation in the study or to withdraw consent to participate at any time without reprisal.

Stimuli. As stimuli, we used the same 16 Japanese calligraphies (see Table 1) as Gershoni and Hochstein (2011). The calligraphies were positioned in a white square of 450 × 450 pixels and presented in the center of the monitor on a gray background. Stimulus presentation and response registration were controlled by SoSci Survey (Leiner, 2014). The device (computer, smartphone, etc.) and screen size at the user frontend were also registered.

Table 1. Mean Ratings (Balance and Liking) for the Japanese Calligraphies in Experiment 1.

Mean balance ratings in ascending order

8	**16**	**13**	**14**	**5**	**6**	**4**	**15**	**1**	**7**	**11**	**12**	**9**	**10**	**3**	**2**
42	47	47	48	48	48	51	59	59	61	63	65	71	78	79	80

Mean liking ratings in ascending order

5	**2**	**8**	**9**	**6**	**14**	**7**	**1**	**12**	**13**	**16**	**11**	**3**	**4**	**10**	**15**
38	39	43	51	52	53	54	56	56	58	60	61	63	64	65	66

Note. Bold numbers represent the picture IDs.

For all stimuli, we calculated objective measures of balance (APB, DCM, and HG scores) and complexity (FS and MS scores). The APB scores ranged from 10.5 to 38.9 ($M = 23.1$, $SD = 7.65$), the DCM scores from 2.51 to 19.6 ($M = 9.57$, $SD = 5.59$), the HG scores from 65 to 87 ($M = 79.8$, $SD = 6.37$), the FS scores from 12 to 36 ($M = 25.6$, $SD = 6.15$), and the MS scores from 1.34 to 19.1 ($M = 11.8$, $SD = 4.67$).

Procedure. The online experiment started with an instruction that informed the participants about the task. In addition, we used a seriousness check (Reips, 2002, 2009) to control for participants' involvement on the task. Subsequently, we used a counterbalanced design with two blocks in each of which the 16 stimuli were presented in random order. Participants had to rate how much they liked the stimuli (1 = *I do not like it* to 6 = *I like it*) in one block and how well the stimuli were balanced (1 = *not balanced* to 6 = *balanced*) in the other block. The ratings were entered by clicking on one of the six scale positions with the mouse. Shortly after the response, the next stimulus was displayed. There was no time limit. In total, the experiment lasted about 5 minutes.

Results

All 149 participants indicated in the seriousness check that they wanted to participate seriously; 40.3% conducted the online experiment on a computer (screen width ranged from 1,138 to 2,048 pixels, and screen height ranged from 640 and 1,152 pixels) and 59.7% on a smartphone (screen width between 320 and 412 pixels, and screen height between 568 and 846 pixels). Despite the use of different devices with unequal screen sizes, ratings of computer and smartphone users produced similar ratings: balance ratings: $r = .967$, $p < .001$, 95% CI [0.906, 0.989]; and liking ratings: $r = .918$, $p < .001$, 95% CI [0.776, 0.972] (correlations of mean values of every single stimuli across participants). Thus, our data did not depend on the use of a specific device.

For comparison with the original study, we also multiplied the ratings by 16, in order to adjust the values to the range of the balance scores (0 to 100). We then computed the mean balance and liking ratings across participants for each of the 16 stimuli. The correlation between balance and liking ratings was not reliable ($r = .374$, $p = .154$, 95% CI [–0.149, 0.734]). To test whether there were sequential effects between the rating types, we also analyzed the between-participants data separately for each block of ratings. For the first block, the correlation between balance (B_1) and liking (L_1) across pictures was not significant ($r = .168$, $p = .533$, 95% CI [–0.357, 0.613]). However, for the second block, after which the participants had already assessed balance or liking, respectively, the correlation (between B_2 and L_2) was significant ($r = .529$, $p = .035$, 95% CI [0.044, 0.812]). This increased correlation was due to influences from picture rating in Block 1 to that in Block 2. When we used partial correlation to control for the variances of the ratings from Block 1, then the correlation between the rating in Block 2 changed to $r = -.170$, $p = .545$. This confirms the substantial interference between the blocks. Thus, to avoid potential biases and misinterpretations, we used only the between-participants data for further analyses.

For these data, the mean balance and liking ratings were 59.1 ($SD = 12.7$) and 54.9 ($SD = 8.76$), respectively. The mean ratings for each stimulus are shown in Table 1. Importantly, our balance ratings correlated highly ($r = .943$, $p < .001$, 95% CI [0.840, 0.980]) with those observed by Gershoni and Hochstein (2011).

The correlations between our mean ratings and the considered scores of balance and complexity are shown in Table 4. As can be seen, there was no reliable correlation ($r = .071$, $p = .793$, 95% CI [–0.440, 0.548]) between perceptual balance and the APB scores, which replicates the results from Gershoni and Hochstein (2011). Moreover, the absent correlation also held for the DCM and HG scores. Merely the complexity measure FS correlated significantly with the balance ratings ($r = -.573$, $p = .020$, 95% CI [–0.832, –0.108]). Finally, the relation between balance ratings and liking ratings, which is also shown graphically in Figure 1, was not significant ($r = .168$, $p = .533$, 95% CI [–0.357, 0.613]).

By averaging the ratings for each picture across participants, much information is lost. Therefore, to increase statistical power, we also analyzed the data by applying linear mixed-effects models (LMMs). These models use all data from each participant and allow to take interindividual differences and stimulus-specific effects into account by treating participants and stimuli as random factors.

The LMMs were computed with R (R Core Team, 2017) and the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). We analyzed the extent to which the formal scores (fixed effects, without interaction term) account for the liking and balance ratings, respectively. Before examining the specific associations, we checked which model specification contributes significantly to the model's goodness of fit (Baayen, Davidson, & Bates, 2008). Therefore, for every association, we conducted a likelihood ratio test that compared the version of the model in which intercepts were allowed to vary across participants and stimuli (random-intercept model), with the version in which participants and stimuli were also allowed to have different slopes (random-slope model). For most of the associations, the random-slope model revealed no significant improvement in the model's goodness of fit (p-values $> .05$). Exceptions were the models with liking ratings as criterion and HG, MS, and FS as predictors. For these models, assuming varying slopes seemed justified. For the remaining models, we assumed only random intercepts for participants and stimuli. In addition, all predictors were used after a grand-mean-centering. Moreover, we applied the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) to calculate the degrees of freedom using the Satterthwaite's approximations for the t test and corresponding p values. The results, except the variance components of participant and stimulus random slopes, because they were smaller than 1%, are

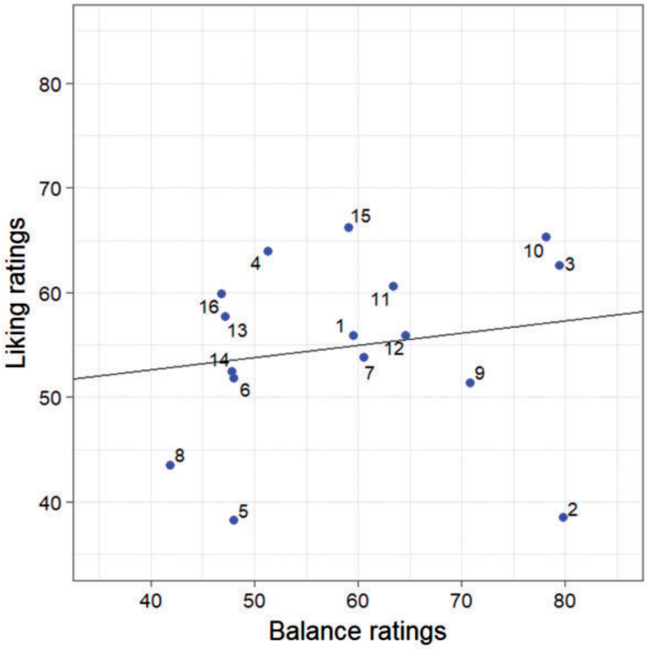

Figure 1. The relation between the mean balance ratings and mean liking ratings for the 16 calligraphies (see Table 1).

shown in Table 2. Our analyses revealed that no measure predicts balance or liking ratings, except FS, which significantly predicts the balance ratings ($p = .013$).

Discussion

Despite the methodological differences between our and Gershoni and Hochstein's (2011) study, the balance ratings were highly correlated. This shows that balance assessments are largely unaffected by stimulus duration and spatial uncertainty. Accordingly, in our experiment, there was also no significant relation between balance ratings and the APB scores, as in Gershoni and Hochstein, which also holds for the other balance measures (Hübner & Fillinger, 2016). Merely complexity (FS) was reliably correlated with the balance ratings. Because the correlation was negative, this means that the less complex the picture (smaller file size), the more it was perceived as balanced.

Thus, our data replicate and generalize Gershoni and Hochstein's (2011) result that the theoretical concepts of balance, as reflected by the considered computational measures, do not correspond to the balance perceived for the Japanese calligraphies. This suggests that perceptual balance is not a unique

Table 2. Results of the LMMs with the Objective Measures as Fixed Effect, and Participants and Stimuli as Random Effects.

Predictor	Criterion	Coefficient (*SE*)	*df*	*t*-value	*p*	σ^2_P (%)	σ^2_S (%)
ABP	Balance	0.118 (0.415)	16.0	0.285	.779	12.0	25.7
	Liking	0.025 (0.288)	16.0	0.088	.931	10.8	11.4
DCM	Balance	0.086 (0.570)	16.0	0.152	.881	12.0	25.8
	Liking	–0.436 (0.379)	16.0	–1.15	.266	10.9	10.6
HG	Balance	–0.592 (0.454)	16.0	–1.31	.210	12.3	23.9
	Liking	1.28 (0.921)	12.9	1.39	.189	11.3	12.1
MS	Balance	0.330 (0.677)	16.0	–0.488	.632	12.0	25.5
	Liking	0.442 (0.558)	11.9	0.792	.444	11.9	6.64
FS	Balance	–1.18 (0.424)	16.0	–2.78	.013	13.1	18.8
	Liking	0.075 (0.479)	7.30	0.157	.879	12.6	1.88

Note. ABP = Assessment of Preference for Balance; DCM = Deviation of the Center of "Mass"; HG = homogeneity; MS = mirror symmetry; FS = file size; *SE* = Standard error; *df* = degrees of freedom; σ^2 = variance components of participant (P) and stimulus (S) random intercepts.

concept. Rather, it seems that for different picture types, observers apply different concepts of balance, which are based on different features or feature combinations. One possible alternative concept will be considered in the next experiment.

An unexpected result was that the balance ratings were not related to the liking ratings. Several reasons are conceivable why this was the case. For instance, it is possible that only specific concepts of balance are related to aesthetic appreciation and that the concept used for assessing the calligraphies is not among them. However, there could also have been methodological reasons. In the studies of Wilson and Chatterjee (2005) and Hübner and Fillinger (2016), simple stimuli were used that were specifically constructed to vary mainly in the balance dimension. In contrast, the real calligraphies used here and in Gershoni and Hochstein (2011) vary also in other dimensions. Thus, even if balance affected liking, its effect might have remained undetected, because it was masked by other dimensions that affected liking to a much larger extent. If we would know these dimensions, then we could control for their contribution to liking and test, whether the remaining variance correlates significantly with balance. In the next experiment, we examined such a dimension.

Experiment 2

The goal of our second experiment was to further investigate the concept of perceptual balance for Japanese calligraphies and its relation to liking. As we have seen in Experiment 1, the balance ratings were highly correlated with those observed by Gershoni and Hochstein (2011). This indicates that persons agree on what "balance" means for these stimuli. However, the fact that the ratings did not correlate with the formal measures of balance suggests that balance has a specific meaning for Japanese calligraphies that is not reflected by these measures. To get an idea which feature could have determined balance perception, we inspected the calligraphies in Table 1 and their order with respect to the balance ratings. It seemed to us that the stimuli assessed as less balanced show some kind of instability and flexibility, whereas calligraphies rated as more balanced look more stable and rigid. In art theory, stability is considered as a visual and aesthetic habit that plays an important role in composing pictures (Liu et al., 2017). This habit is formed in individuals through their lifetime by living with gravitation, which forces them to arrange things in a stable way. In this respect, others also speak of "gravitational stability" (van der Helm, 2015). To examine whether visual balance actually means perceived stability for the Japanese calligraphies, we asked our participants to assess the stability of these stimuli.

Another aim of the present experiment was to find a variable that strongly determined the liking ratings of the calligraphies in Experiment 1. Such a variable would allow us to control its effect and test whether the unexplained variance can be accounted for by balance. However, which property of the calligraphies strongly determines how much they are liked? After inspecting the pictures and their order with respect to liking (Table 1), we hypothesized that prototypicality (Rosch, 1975) could be related to liking. It has been shown before that prototypicality can have a positive effect on preference for design (Hekkert et al., 2003; Whitfield & Slatter, 1979). Furthermore, this property is related to perceptual fluency, which also affects liking (e.g., Graf & Landwehr, 2017). Therefore, we asked our participants to also assess the prototypicality of each stimulus.

Method

Participants and Stimuli. We again used ORSEE (Greiner, 2015) for the recruitment of 152 (39 men, mean age 24.3 years, $SD = 7.65$) participants. As incentive, participants had the chance to win a voucher for their participation. The experiment was performed under the same ethical standards as the previous experiment. We used the same stimuli as in Experiment 1. Moreover, stimuli were presented online with SoSci Survey (Leiner, 2014) under same conditions as in the previous experiment.

Procedure. The procedure was similar to that in Experiment 1. We used a counterbalanced block design. In one block, participants had to rate prototypicality of the stimuli, that is, we asked "Is this a typical Asian character?" (1 = *completely untypical* to 6 = *very typical*). In addition, in the other block, we asked the participants how stable the arrangement of pictorial elements appears ($1 = $ *instable* to $6 = $ *stable*). Participants entered their judgment by clicking on a 6-point scale. There was no time limit. In total, the experiment lasted about 5 minutes.

Results

Two of the 152 participants indicated in the seriousness check that they did not want to participate seriously. Consequently, the corresponding data were excluded from analysis. Furthermore, 49.3% of the participants used a computer (screen width ranged from 1,280 to 2,560 pixels, and screen height ranged from 720 and 1,440 pixels) for the task, 48% a smartphone (screen width between 320 and 732 pixels, and screen height between 412 and 1,107 pixels), and 2.0% a tablet (screen width between 601 and 768 pixels, and screen height between 962 and 1,024 pixels). One persons' device was not identifiable. Importantly, regardless of the used device and screen size, the performance showed a high consistency: prototypicality ratings: $r = .992$, $p < .001$, 95% CI [0.977, 0.997]; and stability ratings: $r = .980$, $p < .001$, 95% CI [0.942, 0.993] (correlations of mean values of every single stimuli across participants).

We first computed the mean prototypicality and stability ratings (multiplied by 16) across participants for each of the 16 stimuli. There was a reliable correlation between these two variables ($r = .688$, $p = .003$, 95% CI [0.292, 0.883]). However, when we analyzed the between-participants data for the first and second block separately, then the correlation was smaller for the first block ($r = .467$, $p = .068$, 95% CI [–0.037, 0.782]) than for the second one ($r = .818$, $p < .001$, 95% CI [0.543, 0.934]). This indicates that the ratings in the second block were again strongly influenced by the ratings in the first block. When we controlled for the ratings in Block 1, the correlation between prototypicality and stability decreased to $r = –.163$, $p = .562$. Therefore, we used only the ratings from the first block (between participants) for the further analyses.

Means of prototypicality and stability ratings were 61.4 ($SD = 22.3$) and 59.0 ($SD = 14.2$), respectively. The means for each stimulus are shown in Table 3.

As can be seen in Table 4, prototypicality strongly correlated with the liking ratings from Experiment 1 (see also Figure 2). Moreover, there was a reliable correlation of prototypicality with the balance ratings from the previous experiment. As expected, the stability ratings also correlated highly with the balance ratings. However, stability did not correlate significantly with liking. Finally, there were no reliable correlations between the present ratings and the objective measures of balance and complexity.

Analogous to Experiment 1, LMMs were computed with typicality or stability ratings as dependent variable and the formal measures as predictors. The analyses revealed that no measure was able to predict prototypicality or stability ratings.

Table 3. Mean Prototypicality and Stability Ratings of the Japanese Calligraphies in Experiment 2.

Mean prototypicality ratings in ascending order

5	**13**	**7**	**14**	**8**	**6**	**2**	**12**	**1**	**9**	**16**	**3**	**11**	**4**	**15**	**10**
27	27	37	40	46	53	57	59	62	66	67	82	87	90	91	91

Mean stability ratings in ascending order

13	**6**	**1**	**12**	**16**	**14**	**7**	**4**	**11**	**8**	**5**	**9**	**15**	**3**	**10**	**2**
33	39	45	51	51	54	54	57	61	63	63	65	68	74	82	85

Note. Bold numbers represent the picture IDs.

Table 4. Correlations (Across Stimuli) Between the Mean Balance, Liking (Experiment 1), Prototypicality, and Stability Ratings (Experiment 2) and Objective Measures for Balance and Complexity.

	Liking	Prototypicality	Stability	APB	DCM	HG	MS	FS
Balance	.168	.498*	.699**	.071	.038	–.311	–.121	–.573*
Liking	–	.700**	–.088	.022	–.278	.268	.047	.095
Prototypicality	–	–	.467	.232	.072	.044	.122	–.201
Stability	–	–	–	.106	.213	–.127	.220	–.353

Note. ABP = Assessment of Preference for Balance; DCM = Deviation of the Center of "Mass"; HG = homogeneity; MS = mirror symmetry; FS = file size.
* $p < .05$. ** $p < .01$.

The relationships between liking and the other ratings were further analyzed by hierarchical linear regressions with liking as dependent variable. To control for the effect of prototypicality, this variable was always entered first as predictor ($R^2 = .490$, $F(1, 14) = 13.5$, $p = .003$), followed by the variable of interest in conjunction with the two-way interaction between prototypicality and the variable of interest. In addition, we separately analyzed the contributions of the single terms of the conjunction (variable of interest and interaction). When we did the analysis with stability, the conjunction of the predictor and the two-way interaction accounts for 38% of the variance, $\Delta R^2 = .377$, $\Delta F = 17.1$, $p < .001$. Furthermore, the separate analysis of the conjunctions' components showed that stability accounts for about 22% of the variance, $\Delta R^2 = .221$, $\Delta F = 9.93$, $p = .008$, and, moreover, the two-way interaction explains 16% of the variation in liking, $\Delta R^2 = .157$, $\Delta F = 14.2$, $p = .003$. This significant interaction indicates that the effect of stability on liking depends on the prototypicality of the stimuli.

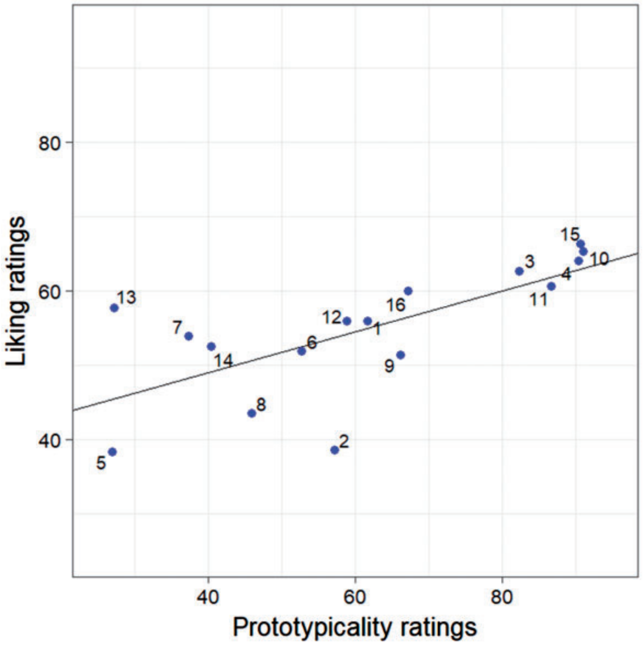

Figure 2. The relation between the mean prototypicality (Experiment 2) and mean liking ratings (Experiment 1) for 16 calligraphies (see Table 3).

To illustrate this interaction graphically, we used a median split to divide the calligraphies into two categories of more and less typical exemplars, respectively. We then computed the correlations between liking and stability separately for the two categories and plotted the data of each category together with their respective regression lines. As can be seen in Figure 3, if prototypicality was low, stability had a negative effect on liking ($r = -.829$, $p = .011$, 95% CI [–0.968, –0.299]). However, if it was high, stability had no significant effect on liking ($r = .462$, $p = .250$, 95% CI [–0.360, 0.880]).

In a further analogous regression analysis, we substituted stability by perceptual balance (Experiment 1). The overall effect of balance and its two-way interaction with prototypicality did not contribute significantly to liking, $\Delta R^2 = .072$, $\Delta F = 0.989$, $p = .400$, which also held for the individual components (balance: $\Delta R^2 = .043$, $\Delta F = 1.21$, $p = .291$; and Balance × Prototypicality: $\Delta R^2 = .029$, $\Delta F = 0.787$, $p = .392$).

With the same schema of hierarchical regression, we tested whether the objective measures predict liking when controlled for prototypicality. For the measures APB, HG, MS, and FS, there were no effects. However, the DCM together with its interaction with prototypicality significantly predicts liking, $\Delta R^2 = .206$, $\Delta F = 4.05$, $p = .045$. To illustrate this interactive relation graphically, we

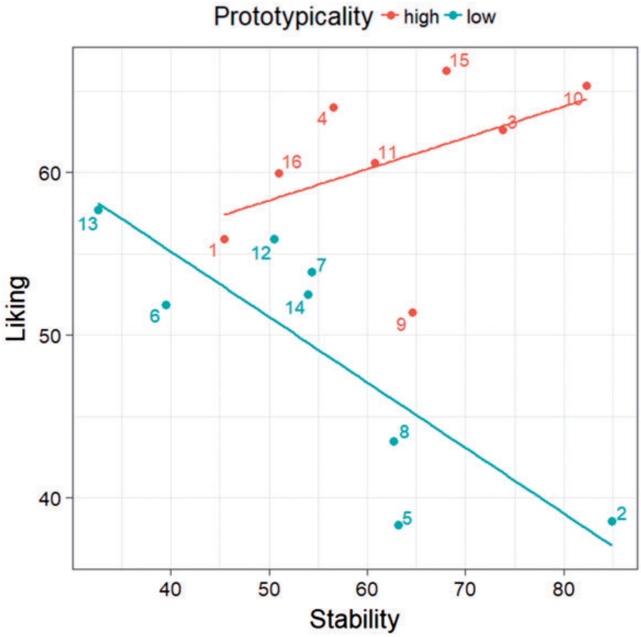

Figure 3. The interaction between prototypicality and stability for the prediction of liking (see Table 3).

computed the correlation between DCM and liking separately for typical and atypical calligraphies. As can be seen in Figure 4, for atypical calligraphies liking substantially increased with a decreasing DCM ($r = -.861$, $p = .006$, 95% CI [–0.974, –0.397]), whereas no significant relation was present for typical calligraphies ($r = -.550$, $p = .158$, 95% CI [–0.904, 0.252]).

Discussion

In this experiment, we collected ratings of stability (Liu et al., 2017; van der Helm, 2015) as well as of prototypicality (Rosch, 1975) for the Japanese calligraphies already used in our first experiment. Stability ratings were used to examine to what extent this dimension corresponds to the concept of perceptual balance for the calligraphies. Indeed, stability ratings correlated highly with the balance ratings from Experiment 1, which indicates that the two concepts are closely related. Accordingly, stability did not correlate with liking, similar to perceptual balance. However, this only held for their direct linear relation.

As expected, prototypicality ratings were highly correlated with the liking ratings from Experiment 1, which supports the idea that people prefer typical members of a category to less typical ones (Hekkert et al., 2003; Whitfield &

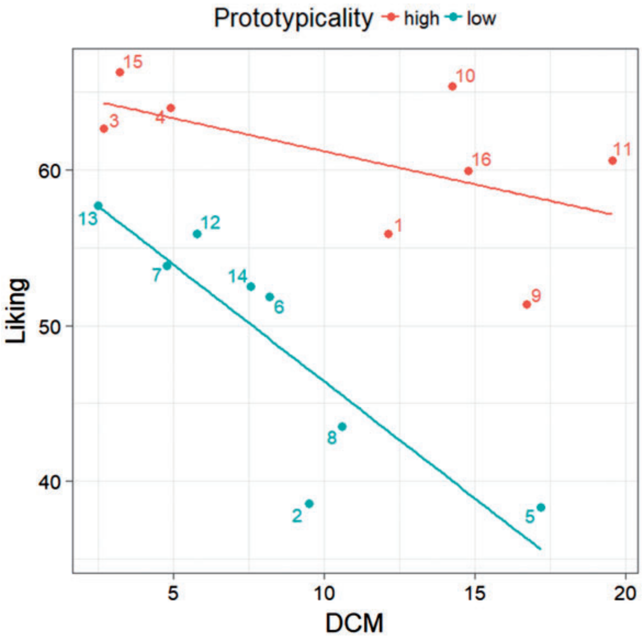

Figure 4. The interaction between prototypicality and DCM for the prediction of liking (see Table 3).

Slatter, 1979). Most importantly, though, this relation allowed us by means of hierarchical multiple regression to discount the effect of prototypicality on the liking ratings and to examine to what extent the remaining variance can be explained by the other variables. When we did this for stability, it turned out that the variable had a significant effect on liking. Moreover, there was also a reliable two-way interaction between prototypicality and stability indicating that stability had a negative effect on liking, but only for less typical calligraphies. For more typical exemplars, there was a trend in the opposite direction (see Figure 3).

For the balance ratings from our first experiment, an analogues hierarchical regression analysis revealed no significant effects on liking, even when the effect of prototypicality was controlled. This shows that the perception of stability and balance are similar but not identical, which indicates that they are based, at least partly, on different visual features.

For the different formal measures, hierarchical regression revealed no reliable relations with liking, except for the DCM, which in combination with its interaction with prototypicality significantly predicted liking. Further analyses revealed that for less typical calligraphies, liking decreased with an increasing DCM, that is, with an increasing distance of the center of mass from the geometrical center of the picture. This relation was much weaker and unreliable for prototypical calligraphies (see Figure 4).

General Discussion

In this study, we investigated the proposed positive relation between the perceptual balance of a picture and its aesthetic appreciation (Arnheim, 1982; Bouleau, 1980; Kandinsky, 1926/1979). Up to now, this hypothesis has mainly been supported by a study with photographs (McManus et al., 2011), and a few studies with simple and specifically constructed stimuli (Hübner & Fillinger, 2016; Wilson & Chatterjee, 2005). In the latter studies, it has also been shown that computational measures of balance can be constructed that predict not only balance but also liking ratings. Gershoni and Hochstein (2011), however, applied one of such measures (APB) to Japanese calligraphies and failed to find a relation with balance ratings. This is an important result because it seems to demonstrate possible limits of the balance measures. Therefore, the first aim of this study was to replicate and extend Gershoni and Hochstein's result. Specifically, because these researchers presented their stimuli for a relatively short time and with spatial uncertainty, we investigated in our first experiment whether the results also occur with a longer stimulus presentation time and without spatial uncertainty. Furthermore, we wanted to apply different measures of balance as well as measures related to visual complexity. In addition, we collected liking ratings to examine the relation between balance and aesthetic appreciation for Japanese calligraphies. Because Gershoni and Hochstein (2011) used only 16 calligraphies, the statistical power of their tests was relatively small. Nonetheless, we used the same small set to ensure comparability between the original and our study.

As the results show, our balance ratings correlated highly with those from the original study, which demonstrates that balance perception is rather robust against procedural variations. Moreover, the balance ratings did not correlate with any of the balance measures. This indicates that the balance perceived when looking at calligraphies is unrelated to the concept of balance reflected by the applied formal measures. Finally and unexpectedly, our balance ratings, as well as the formal measures, did also not correlate with the liking ratings.

The results of our first experiment raised at least two questions. First, what does the concept of balance as assessed by our participants mean for calligraphies? Second, why was there no relation between balance and liking? For answering the first question, we inspected the order of pictures with respect to balance. Our impression was that perceived stability (Liu et al., 2017; van der Helm, 2015) might explain some of the variance. Therefore, in our second experiment, we asked our participants to assess the stability of the calligraphies.

As result, the stability ratings correlated highly with the balance ratings from Experiment 1. This suggested that, instead of assessing how well the picture elements are balanced, the participants judged, at least to a large extent, how stable or flexible the calligraphies look like. Given this specific concept of balance, it is not surprising that the formal measures did not correlate with the ratings. Because stability did also not correlate with liking, this concept of balance seemed to be unrelated to aesthetic appreciation. However, these conclusions are premature. With respect to the absent correlation between balance and liking, we also hypothesized that a variable other than balance might strongly determine the aesthetic appreciation of Japanese calligraphies, which makes it difficult to detect a presumably relatively small effect of balance. As likely candidate, we considered prototypicality, which has shown to be related to aesthetic preference (Hekkert et al., 2003; Whitfield & Slatter, 1979). Therefore, in our second experiment, we additionally collected ratings for this property.

As expected, prototypicality was highly correlated with liking. This allowed us by means of hierarchical multiple regression to control the effects of prototypicality on liking. When we did this with respect to stability, we found a reliable relation between stability and liking. However, there was also a significant interaction between stability and prototypicality. Stability was related to liking only for stimuli rated as more atypical for Japanese calligraphies but not for more typical ones. Interestingly, an analogous analysis of the balance ratings from Experiment 1 revealed no reliable relation with liking. This indicates that despite their substantial correlation, stability and balance ratings also differ in some relevant aspect.

Discounting the effects of prototypicality also revealed an interesting relationship between the DCM and the liking ratings observed in Experiment 1. It turned out that the DCM scores affected only the liking ratings for more atypical members of the calligraphies.

Thus, the present approach demonstrates that it can be difficult to detect effects of a certain variable on liking ratings if the considered pictures not only vary on the corresponding dimension but also on other dimensions affecting aesthetic appreciation. In this case, the dimensions can compete and interact in a complex way, a phenomenon called “gestalt nightmare” by Makin (2017). However, in case such competing dimensions are known, their effects can be discounted, which allows analyzing the relation between the unexplained variance and the variable of interest. Thus, our results provide a positive example where the gestalt nightmare could be avoided.

Our conclusions are based on the between-participants data only to avoid influences of sequential effects between the two blocks of ratings. This reduced the statistical power of our tests. However, because the sequential effects were statistically significant, discarding the data of the respective second block was the only way to prevent biases and misinterpretations. Beyond that, the sequential interferences between the different ratings is an important observation that should be taken into account when designing future studies in this area.

Taken together, the results of the present study show that the meaning of balance can differ between different stimulus types. For the Japanese calligraphies used here, it seems that balance had a different meaning than the usual interpretation in art theory (Arnheim, 1982; Bouleau, 1980; Kandinsky, 1926/1979). It was neither related to the considered formal measures nor to the liking ratings. However, the related concept of stability was predictive for liking, even if only for the more atypical members of the calligraphies. For more prototypical members, stability had no effect on aesthetic appreciation. In future research, the concept of stability and its relation to other concepts of balance and to aesthetic appreciation should be investigated further. For instance, it would be interesting to know to which picture types the concept of visual stability applies. Until now, it is not clear which kind of pictorial arrangement is needed to evoke the perception of instability. Our results suggest that stability perception comes into play when different figural elements in a picture are connected rather than disconnected.

The fact that for more atypical calligraphies the DCM was related to liking suggests that this measure reflects a relatively general formal concept of balance, although it is obviously not always taken into account for the assessment of balance.

Authors' Note

The full data set for this article is accessible on https://osf.io/urmbk/.

Acknowledgments

The authors thank Sebastian Dummel for his methodical advice regarding the linear mixed-effects models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Martin G. Fillinger http://orcid.org/0000-0001-9532-1699

References

Arnheim, R. (1954). *Art and visual perception: A psychology of the creative eye*. Berkeley: University of California Press.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bouleau, C. (1980). *The painter's secret geometry: A study of composition in art*. New York, NY: Hacker Art Books.

Donderi, D. C. (2006). An information theory analysis of visual complexity and dissimilarity. *Perception*, *35*(6), 823–835. doi:10.1068/p5249

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., & Sawey, M. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology*, *102*(1), 49–70. doi:10.1348/000712610X498958

Gartus, A., & Leder, H. (2017). Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception. *PLoS One*, *12*(11), e0185276. doi:10.1371/journal.pone.0185276

Gershoni, S., & Hochstein, S. (2011). Measuring pictorial balance perception at first glance using Japanese calligraphy. *i-Perception*, *2*(6), 508–527. doi:10.1068/i0472aap

Graf, L. K. M., & Landwehr, J. R. (2017). Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Frontiers in Psychology*, *8*, 15. doi:10.3389/fpsyg.2017.00015

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. doi:10.1007/s40881-015-0004-4

Hekkert, P., Snelders, D., & Van Wieringen, P. C. W. (2003). Most advanced, yet acceptable: Typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology*, *94*(1), 111–124. doi:10.1348/000712603762842147

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

Jacobsen, T. (2004). Individual and group modelling of aesthetic judgment strategies. *British Journal of Psychology*, *95*(1), 41–56.

Kandinsky, W. (1926/1979). *Point and line to plane*. New York, NY: Dover.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 26. doi:10.18637/jss.v082.i13

Leiner, D. J. (2014). SoSci Survey (Version 2.5.00). Retrieved from https://www.soscisurvey.de

Liu, J., Dong, W., Zhang, X., & Jiang, Z. (2017). Orientation judgment for abstract paintings. *Multimedia Tools and Applications*, *76*(1), 1017–1036. doi:10.1007/s11042-015-3104-5

Locher, P. J., Krupinski, E. A., Mello-Thoms, C., & Nodine, C. F. (2007). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, *21*(1/2), 55–77. doi:10.1163/156856808782713762

Locher, P. J., & Nagy, Y. (1996). Vision spontaneously establishes the percept of pictorial balance. *Empirical Studies of the Arts*, *14*(1), 17–31.

Locher, P. J., & Stappers, P. J. (2002). Factors contributing to the implicit dynamic quality of static abstract designs. *Perception*, *31*(9), 1093–1107. doi:10.1068/p3299

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., & Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, *160*, 43–57. doi:10.1016/j.actpsy.2015.06.005

Makin, A. D. J. (2017). The gap between aesthetic science and aesthetic experience. *Journal of Consciousness Studies*, *24*(1-2), 184–213.

McManus, I., Edmondson, D., & Rodger, J. (1985). Balance in pictures. *British Journal of Psychology*, *76*(3), 311–324. doi:10.1111/j.2044-8295.1985.tb01955.x

McManus, I., 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*(6), 615–647. doi:10.1068/i0445aap

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

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. doi:10.1026//1618-3169.49.4.243

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. *Human Vision and Electronic Imaging XIV, Proceedings of SPIE*, *7240*, 724008.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. doi:10.1037/0096-3445.104.3.192

van der Helm, P. A. (2015). Symmetry perception. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization* (pp. 108–128). Oxford, England: Oxford University Press.

Whitfield, T. W. A., & Slatter, P. E. (1979). The effects of categorization and prototypicality on aesthetic choice in a furniture selection task. *British Journal of Psychology*, *70*(1), 65–75. doi:10.1111/j.2044-8295.1979.tb02144.x

Wilson, A., & Chatterjee, A. (2005). The assessment of preference for balance: Introducing a new test. *Empirical Studies of the Arts*, *23*(2), 165–180.

Author Biographies

Martin G. Fillinger is a research associate at the Department of Psychology, Universität Konstanz. His work focuses on the connection between visual balance and aesthetic appreciation.

Ronald Hübner is a professor of Cognitive Psychology at the Department of Psychology, Universität Konstanz. His work focuses on empirical aesthetics and visual attention.