Selective social influence on aesthetic evaluations via natural language testimony

Yoko Urano (yokourano@berkeley.edu)
Department of Psychology, University of California, Berkeley

Bill Thompson (wdt@berkeley.edu)
Department of Psychology, University of California, Berkeley

Abstract

Why and how do we incorporate others' judgments when making an aesthetic evaluation? We investigated this question by studying social transmission of aesthetic evaluations via natural language, which conveys richness that more common (numerical) measures may fail to convey, such as the reasoning behind a judgment. Participants in a large-scale study aesthetically evaluated photographs, either independently or after observing testimony from another person. We found that participants formed more similar evaluations to the testimony they observed (than the asocial control). Furthermore, participants who received the same evaluative testimony wrote evaluations that were more similar to each other in *content* but not sentiment (relative to a matched asocial cohort). This suggests that social influence on aesthetic evaluations may have a greater informational aspect than previously understood.

Keywords: social influence; aesthetics; natural language processing

Introduction

Walking through an art museum with a friend affords the eye-opening experience of 'seeing what they see'; they may point out features of a painting that you had not noticed or express distaste for a sculpture that you thought was actually quite nice. Oftentimes, the information you get from this friend can shift how you judge the work—maybe that sculpture *is* uglier than you thought at first glance. While aesthetic evaluations are often highly personal and idiosyncratic, we also tend to care about how others evaluate the same object. Why and how do we incorporate others' opinions when forming an aesthetic judgment? What social dynamics explain why we would not disregard someone else's opinion entirely?

There are at least two differnt mechanisms through which social influence may apply to aesthetic judgment (Deutsch & Gerard, 1955). The first mechanism is *Normative influence*. This perspective supposes that aesthetic judgments are a means for social cohesion via reinforcement of norms. This type of influence has been well-documented in empirical studies; normative signals such as aggregate statistics and expert opinion have been shown to influence aesthetic judgments towards that signal (e.g., Urano, Marjieh, Griffiths, & Jacoby, 2024; Hullman, Adar, & Shah, 2011; Altenmüller & Plewe, 2024; Salganik, Dodds, & Watts, 2006; Hesslinger, Carbon, & Hecht, 2017).

The second mechanism is *Informational influence*. This perspective on aesthetics places a greater emphasis on our desire to have a precise and refined aesthetic judgment. While it has received less empirical investigation, its relevance is implied by the philosophical notion of aesthetic authenticity. Aesthetic authenticity posits that our judgments should be reflective of our authentic selves. Under this perspective, social influence on aesthetic judgement may initially seem surprising: we might hestitate to take someone else's aesthetic testimony at face value, even when we know them to be honest and competent (Bräuer, 2023). However, authenticity suggests that we might instead consider the *reasons* other people give for their evaluation, and use this information to derive our own. The measures used in aesthetics research in the past—low-dimensional measures such as likert-scale ratings, rank order, and two-alternative forced choice paradigms (Palmer, Schloss, & Sammartino, 2013)—may have underestimated the potential for informational influence because they may not support the transmission of important relevant features, such as the structure of evaluations and the reasoning behind them.

One way to create the potential for both kinds of influence in an experimental study is to study natural language evaluations. Natural language evaluations offer a rich, multidimensional modality that can provide more information than a single number, allowing for both a judgment *and* the evidence it arises from to be transmitted to another person. However, it has traditionally been a challenge to reconcile the complexity of a free-form aesthetic evaluation with the need for well-defined quantitative measures.

Recent advances in natural language processing and language modelling offer solutions to this dilemna. These techniques transform natural language into high-dimensional numerical vectors with which to conduct quantitative analyses. In our study of social influence, participants were exposed to natural language evaluations as social information (i.e. as testimony) and produced natural language evaluations themselves. Not only does this create much richer social information that offers the opportunitity for both normative and informational influence, but it also means that we can analyze responses to testimony in a way that may separate the two. Finally, evaluations expressed in language also relate more directly to naturalistic scenarios where we socially express or are exposed to aesthetic evaluations; how is it that our friend in the art museum, through their *words*, shapes our own judgment? More generally, our approach may be applicable to the study of testimony in other domains such as learning (e.g.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

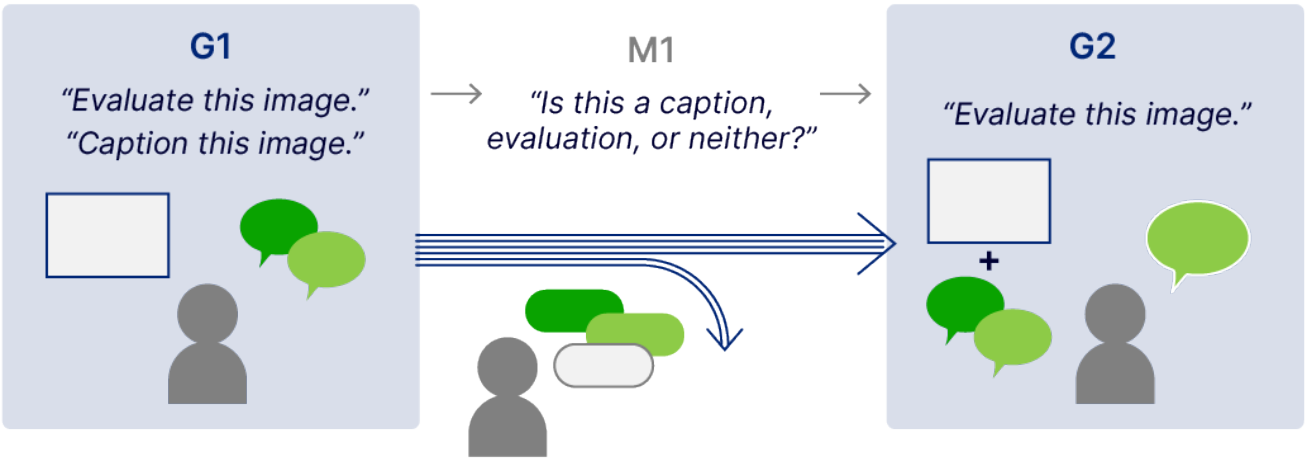

Figure 1: **Evaluations and captions collected by human participants are given as social information to a subsequent generation of participants.** Testimonies in the form of aesthetic evaluations and captions were collected for 50 naturalistic photographs (G1). After a quality filter by an independent group of participants (M1), 30 stimuli were presented along with testimonies to a new generation of participants (G2).

learning about research best practices from an advisor), consumer behavior (e.g. choosing a car in light of reviews), and social perception (e.g. assuming somebody is mean based on hearsay).

In this study, we collected written testimony from human participants about visual stimuli and showed them to a separate group of participants while they write their own evaluations. We measure social influence on these resulting evaluations by calculating their similarity to the given testimony and to each other. In addition to a broad effect of testimony, we find differences in how specific aspects of participant's responses are influenced.

Methods

Participants

We recruited all participants online through Prolific (N=552 across the entire experiment). All participants were 18 or older, fluent in English, and had normal or corrected-to-normal vision. All participants were paid at a rate of 18 USD per hour based on our estimated completion times, which ranged from 8 to 15 minutes.

Materials

We used a subset of photographs from the Impressions dataset (Kruk, Ziems, & Yang, 2023). This subset was selected through a pilot survey measuring ratings of three aesthetic qualities (beauty, interest, and artistic) for each photograph. Our goal was to curate a stimulus set that would induce varied aesthetic responses. We selected the 50 photographs with the greatest sum of variance in rating across these items as our initial stimulus set. This stimulus set was later further subset into 30 stimuli through the quality filtering process described below.

Procedure

Collecting aesthetic testimony Figure 1 depicts an overview of our experimental paradigm. The main experiment consisted of two generations of participants (G1: N=108, G2: N=194). To gather texts that will be given as testimonies in the second generation (G2), participants in the first generation (G1) were each shown 5 randomly selected images from our stimulus set and given two writing prompts for each image: *aesthetic evaluation* and *caption*. *Aesthetic evaluations* were defined to the participants as describing what the image makes one think or feel, especially in regards to how beautiful, interesting, or artistic it is. *Captions* were defined as communicating the visual content of the image (adapted from US government IT accessibility guidelines for alt-text [https://www.section508.gov/create/alternative-text/]). The order of prompts was randomized for each trial. Attention checks were given after the second trial by asking the participants to report the last prompt (i.e., caption or evaluation) they replied to.

Quality control for testimonies To check the quality of responses from G1, we recruited a new group of participants (M1) to judge whether each response is a *caption*, an *aesthetic evaluation*, or *neither* (N=250). These participants were given the same definitions of the prompts as G1 and additionally told that they could select *neither* if unsure. Each (stimulus, text) pair was judged by at least 5 participants. There were 5 repeat trials at the end of the experiment to asses within-participant reliability.

Transmission structure Based on M1 responses, we select (stimulus, testimony) pairs that were judged appropriately as either caption or evaluation by at least two-thirds of the M1 participants who judged that pair. Then, we cre-

ated a subset of these data so that for every stimulus selected, for each caption, we could also use the *evaluation* from the same G1 participant (i.e. both the caption and the evaluation passed the M1 filter). This subset of data includes 180 total (stimulus, testimony) pairs and 30 unique stimuli; for each stimulus, there were 3 captions and 3 evaluations (i.e., caption-evaluation pairs from 3 G1 participants). Participants in our second generation (G2) were then each shown 10 images from this data set and asked to write *aesthetic evaluations* for each one. In the instructions, participants were told that in some cases, they may see what another participant had written (*social condition*); of the 10 images that each participant saw, 5 were accompanied with a response from G1 as testimony and 5 were evaluated without testimony (*asocial condition*). This represents a within-subject manipulation. The order of social and asocial trials was randomized. After writing their evaluations of an image, participants were also asked to rate it on three aesthetic items (how *beautiful*, how *interesting*, and how *artistic* the image is) on a scale of 1 (not at all) to 5 (extremely).

Data Analysis

Quality criteria At each step of our analysis, we excluded participants whose reliability did not meet a certain threshold. For G1 participants, this threshold was whether they passed the attention check, which was to report the last prompt (*evaluation* or *caption*) they had responded to (originally N=108; after exclusion, N=93). For M1 participants, the threshold was whether at least 60% of their responses in the repeat trials were the same as their responses in the corresponding original trials (originally 250; after exclusion, N=190). Finally, for G2 participants, the threshold was that their ratings from the repeats trials and those from the original trials had a correlation of at least .6, or (if they had given the same rating for all repeat trials) if the mean difference between their ratings in the repeat and original trials was less than 1.0 (originally N=194; after exclusion, N=147).

Measures of social influence Figure 2 illustrates the experimental conditions and social influence measures. We examined the pairwise similarity between a social response (i.e., an aesthetic evaluations given in light of another person's aesthetic testimony) to the testimony they received (testimony-response similarity). If social responses were influenced in some way by testimony, the mean pairwise similarity should be significantly different from the mean pairwise similarity between the given testimony and asocial responses (i.e., evaluations made without exposure to others' testimony) for the same stimulus. The latter represents how similar responses are to the given testimony at baseline. If it the influence is attractive, the similarity should be significantly *greater* in the social condition; if the effect is repulsive, we should see the opposite. For each comparison, we also considered whether the type of testimony is an important feature by comparing the impact of testimony as caption and testimony as evaluation. We also calculated the similarity between social and

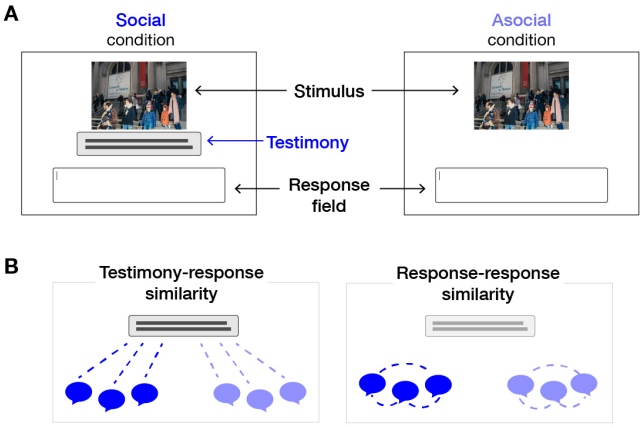

Figure 2: **Overview of experimental conditions (*social* vs *asocial*) and measures of social influence. A.** In each trial, participants were asked to write their aesthetic evaluation of the stimulus in free-form text. The only difference between conditions was the presentation of testimony in the *social* condition. This testimony was written by a participant from a previous sample (G1) and could be one of either: a caption or an aesthetic evaluation. **B.** We operationalized social influence in terms of the similarity between testimony and response (testimony-response similarity) and the similarity between responses to the same (stimulus, testimony) pair (response-response similarity).

asocial responses to the same (stimulus, testimony) pair. The similarity with asocial responses gives us a baseline measure of within-stimulus similarity. If the similarity within social responses is greater than this baseline, we would infer that testimony decreased the diversity of aesthetic evaluations.

Similarity measures We calculated cosine similarities between sentence embeddings extracted from a language model (*all_mpnet_base_v2*). In addition, to address the limitation that cosine similarity is not easily decomposed into intuitive reasons for similarity, we also elicited similarity measures from a frontier large language model (ChatGPT, *gpt-4o-mini*). This allowed us to assess similarity on the basis of specific intuitive features. Our first measure instructed ChatGPT to base its judgement on *sentiment* (positive or negative). Our second measure instructed ChatGPT to base its judgement on *content* (i.e. judge whether two texts are similar based specifically on whether there are shared details mentioned in the text). ChatGPT has been shown to generate similarity judgments that are consistent with human judgments in multiple domains (Marjieh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2024). Given two texts describing the same image, we instructed ChatGPT to generate similarity scores on a scale of 0 (extremely dissimilar) to 10 (extremely similar). Specifically considering the sentiment and content of evaluations will help us understand how testimony as social information impacts not just the overall evaluation but also the evidence

Figure 3: **Examples of stimuli, testimony, and responses.** Social influence was tested on 30 stimuli. Each stimulus was paired with one of 6 possible texts in the social condition—3 captions and 3 evaluations. The top half of the figure shows a stimulus and its caption (testimony). On its right are one example each of a social response (outlined in dark purple) and an asocial response (outlined in light purple). The bottom half of the figure shows a stimulus with one of its evaluations (testimony), with social and asocial responses formatted as described above.

that is used to explain it.

When analyzing each of our similarity measures, we conducted independent t-tests to assess the statistical significance of testimony.

Results

Figure 3 shows examples of the stimuli and participant descriptions and evaluations collected in our experiment.

Social responses are more similar to testimony and to each other.

Figure 4 illustrates the effect of social information on the similarity between responses and social information and to each other. Regardless of the type of testimony, responses given in light of testimony are signficantly more similar to that testimony than they would be at baseline, with a mean cosine similarity of .42 in the social condition and .36 in the asocial condition overall ($p < .001$). Dividing up the data along testimony type shows that this effect is retained for both kinds of testimony; with captions only, the mean cosine similarity is .37 in the social condition and .32 in the asocial ($p < .01$); with evaluations only, the means are .47 (social) and .40 (asocial) ($p < .001$). Social responses are also more similar to each other, with a social mean of .40 and an asocial mean of .38 ($p < .05$). However, this effect is driven by cases where the testimony presented is an evaluation (social mean: .40, asocial mean: .38; $p < .05$), not when it is a caption (social mean: .41, asocial: .39).

Testimony affects both sentiment and content similarities.

To better understand what aspect of responses are being influenced by social information, we compare social and asocial responses in terms of their sentiment and their content. Figure 5 shows that responses made in the social condition are significantly more similar to the testimony both in their sentiment and content, with a mean sentiment similarity of 5.54 for social responses and 4.90 for asocial responses ($p < .001$), and a mean content similarity of 3.84 for social responses and 3.15 for asocial responses ($p < .001$). These effects are retained after separating by type of testimony. With captions, sentiment similarity is significantly higher with social (mean: 4.78) than asocial (mean: 4.18) responses ($p < .001$); content similarity is likewise higher in the social condition with a social mean of 3.49 and an asocial mean of 3.00 ($p < .01$). With evaluations, the mean sentiment similarity to the given testimony is 6.28 for social responses and 5.61 for asocial responses ($p < .05$). Mean content similarity is 4.17 for social responses and 3.29 for asocial ($p < .001$). While effects on testimony-response similarity are broadly uniform across

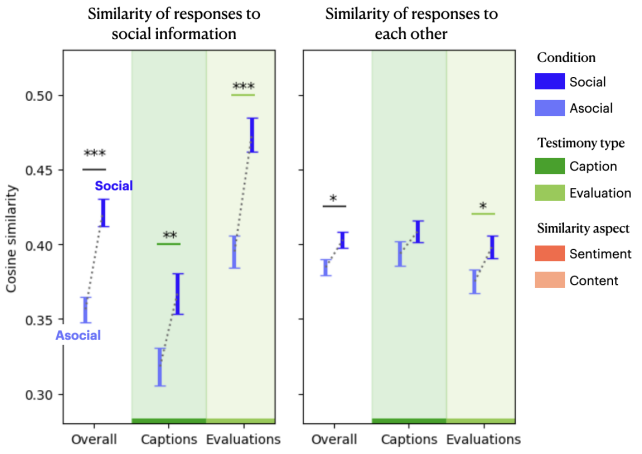

Figure 4: **Testimony significantly affects evaluative responses to stimuli. Left:** Cosine similarities between responses and the social information are significantly greater in the social condition ($p < .001$), suggesting that when people write evaluations in light of social information, they write more similarly to that information than they would at baseline (i.e., asocially). **Right:** Cosine similarities between pairs of responses are also significantly greater overall ($p < .05$), but this effect is driven by evaluations as social information, not captions. Dark blue bars represent the social condition and light blue bars represent the asocial condition. All errorbars represent one standard error.

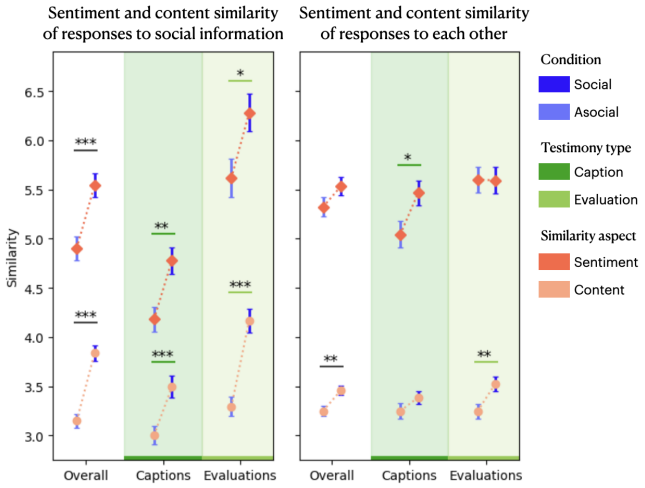

Figure 5: **Testimony affects both sentiment and content of responses in slightly different ways.** Dark orange dots represent sentiment similarity and light orange dots represent content similarity. **Left:** Across both types of testimony, sentiment and content of evaluative responses become significantly more similar to the given testimony ($p < .001$) than they would be at baseline (asocially). **Right:** There is an overall effect of testimony on the content similarity between responses, but not sentiment. When we separate by testimony type, we see that with captions, sentiment becomes more similar between responses but not content; with evaluations, content becomes more similar between responses, with no effect on sentiment.

testimony types and similarity aspects, interesting differences arise when we look at the similarity of responses to each other (rather than to the testimony a person received). Overall, responses in the social condition are significantly more similar to each other in terms of content (social mean: 3.45, asocial mean: 3.25) but not sentiment (social mean: 5.53, asocial mean: 5.32; $p < .01$). Captions increased the similarity of sentiment among responses (social mean: 5.46, asocial mean: 5.04; $p < .05$), but not content (social mean: 3.39, asocial mean: 3.25). Evaluations significantly increased similarity among responses in terms of content (social mean: 3.53, asocial mean: 3.25; $p < .01$), but not sentiment (social mean: 5.59, asocial mean: 5.60). In other words, social responses made after observing evaluations mentioned relatively similar details, but were not particularly similar in terms of the resulting sentiment of their evaluation.

Ratings alone would have failed to identify distinct effects on similarity

Our analysis of similarity between responses revealed unique effects of social influence on the sentiment and content of responses. To test how these effects would appear using a different measurement modality, we conduct an analogous analysis using the ratings we collected in the same experiment. We find that there is no significant difference in the variance of ratings given in the social condition compared to the asocial baseline condition (Figure 6A). In other words, using a low-dimensional measure of aesthetics would not have captured the nuanced effects of testimony as social information.

Captions contain different information than evaluations

We compared the similarity between responses given in light of a caption and the *captioner's own evaluation*, unseen by those respondents, to the same stimulus. This counterfactual transmission test examines whether there is information in the caption that conveys the same information as the evaluation would have. We found that there is no significant increase in the similarity of social responses to the captioner's evaluation, compared with asocial responses (Figure 6B).

Discussion

In our experiment, transmission of natural language aesthetic testimony from person to person influenced aesthetic evaluations. Testimony impacted both sentiment and content. We also found that the kind of testimony matters. When exposed to captions as testimony, responses mentioned similar details to the caption but are not particularly similar among themselves. It is possible that this is because responses each mention only a subset of the details mentioned in the caption, de-

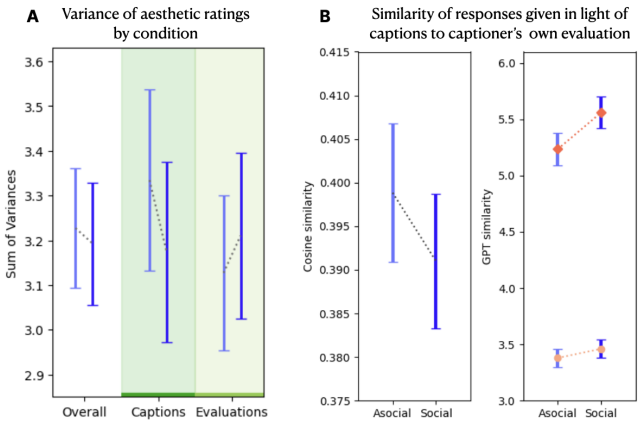

Figure 6: **A.** Ratings do not capture an effect of responses increasing in similarity to each other when given aesthetic testimonies from others. Regardless of the type of information, rating variance does not significantly differ between conditions. **B.** Captions do not increase similarity to the captioner's own evaluations to the same stimulus. **Left:** Cosine similarity does not significantly differ between conditions. **Right:** Dark orange dots represent sentiment similarity and light orange dots represent content similarity. Neither the sentiment nor content of responses are significantly more similar to the captioner's evaluation than they would be asocially.

creasing overlap between them. It is also likely that participants are adding *evaluative* details that differ between them, in addition to the details in the caption. Captions and evaluationd have inherent differeneces in the types of details they are likely to contain; because the participants' task was to write an evaluation, they may have had to add details. This explanation is supported by the fact that these responses are not more similar than asocial responses to the captioner's own evaluation of the same stimulus; captions and evaluations constitute different information. In other words, it seems that responses are drawing different conclusions from the same basic evidence.

When given captions as testimony, responses become more simililar in sentiment to both the caption and to each other. This can be explained as a dampening effect; captions are inherently more neutral than evaluations, so responses given in light of a caption may be becoming more neutral than they would be at baseline. When given evaluations as testimony, responses mention relatively more similar content to both each other and to the testimony. This suggests that they are basing their evaluations on similar details. In contrast to when a caption is given as testimony, there is not as much leeway for participants to add other kinds of details that are not already mentioned in the given testimony, which explains why we see similarity between responses here. When given evaluations as testimony, responses have relatively similar sentiment to the given testimony, but not to each other. One explanation is that they are each being equally influenced in the same direction, as opposed to, say, initially dissimilar evaluations experiencing greater influence than evaluations that were similar from the start.

Taken together, our results suggest that people tend to consider the evidence presented in others' descriptions of a stimulus when forming their own aesthetic evaluations. However, this does not necessarily mean that they draw the same conclusions. In contrast to previous studies on social influence in aesthetic evaluations, our results are more consistent with informational, rather than normative, influence.

It is important to note that social influence in this study lacked certain features of real-world interactions that would constitute normative signals, including exposure to multiple pieces of testimony, the expectation of developing or maintaining a relationship with the source of testimony, and information about a person's aesthetic expertise. We should also consider an alternative explanation of our results, which is that social responses reflect participants' initial judgments (captured as sentiment) that are explained post-hoc using details derived from testimony (i.e., content), akin to the social intuitionist theory of moral judgments (Haidt, 2001). Our results do not disambiguate whether judgment or reasoning occured first; a future study distinguishing this directionality would help us better understand the social influence we see here.

Under the interpretation that we use others' reasoning to draw our own conclusions, we could explain both our results and previous literature by the idea that, in the absence of the reasoning behind another person's evaluation, as is the case in most of previous studies, influence is primarily *normative*. However, once the reasoning becomes available to us, as in our experiment, we may prioritize how we independently feel about those reasons. If our friend in the art museum says, "This sculpture is terrible!", we might be compelled to express a similar judgment in response. If they instead say, "This sculpture is terrible—the proportions are all wrong!", we may consider the proportions, then draw our own conclusion: actually, the sculpture is fashionably *avant-garde*. This is consistent with the notion of aesthetic authenticity and other insights from philosophy that reason that aesthetic judgment comes from our own understanding of the object of evaluation (Bräuer, 2023; Hills, 2022).

This study is a step towards understanding more naturalistic social influence in aesthetic evaluations; it helps to tease apart what drives influence by facilitating transmission of aesthetic evaluations and simple descriptions through natural language, which importantly created the potential to observe both the judgment and the reasoning behind it. We found that social influence in aesthetics may not always be as normative as has previously been suggested. Our study offers empirical data that sheds light on the counterintuitive potential for social influence to meaningfully support aesthetic authenticity.

References

Altenmüller, M. S., & Plewe, M. C. (2024). (Not) alone

in the museum: Implicit social influence on art appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, No Pagination Specified–No Pagination Specified. doi: 10.1037/aca0000713

Bräuer, F. (2023, July). Aesthetic Testimony and Aesthetic Authenticity. *The British Journal of Aesthetics*, *63*(3), 395–416. doi: 10.1093/aesthj/ayac045

Deutsch, M., & Gerard, H. B. (1955, November). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. doi: 10.1037/h0046408

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. doi: 10.1037/0033-295X.108.4.814

Hesslinger, V. M., Carbon, C.-C., & Hecht, H. (2017, December). Social Factors in Aesthetics: Social Conformity Pressure and a Sense of Being Watched Affect Aesthetic Judgments. *i-Perception*, *8*(6), 2041669517736322. (Publisher: SAGE Publications) doi: 10.1177/2041669517736322

Hills, A. (2022). Aesthetic testimony, understanding and virtue. *Noûs*, *56*(1), 21–39. doi: 10.1111/nous.12344

Hullman, J., Adar, E., & Shah, P. (2011, May). The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1461–1470). Vancouver BC Canada: ACM. doi: 10.1145/1978942.1979157

Kruk, J., Ziems, C., & Yang, D. (2023, December). Impressions: Visual Semiotics and Aesthetic Impact Understanding. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12273–12291). Singapore: Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.755

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024, September). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, *14*(1), 21445. doi: 10.1038/s41598-024-72071-1

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013, January). Visual Aesthetics and Human Preference. *Annual Review of Psychology*, *64*(Volume 64, 2013), 77–107. (Publisher: Annual Reviews) doi: 10.1146/annurev-psych-120710-100504

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854-856. doi: 10.1126/science.1121066

Urano, Y., Marjieh, R., Griffiths, T., & Jacoby, N. (2024). The Influence of Social Information and Presentation Interface on Aesthetic Evaluations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).