Your ears don't change what your eyes like: People can independently report the pleasure of music and images

Jessica Frame[1]
Maria Gugliano[2]
Aenne Brielmann[3]
Amy M. Belfi[1]

[1]Department of Psychological Science, Missouri University of Science and Technology
[2]College of Pharmacy, University of Texas at Austin
[3]Department of Computational Neuroscience, Max-Planck Institute for Biological Cybernetics

Author Note: We would like to thank Amanda Lim and Mikayla Shy for their assistance with conducting literature reviews and setting up the experiment in Gorilla. Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number R15AG075609. Aenne Brielmann was supported by the Max Planck Society, the Alexander von Humboldt foundation, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 461354985. Raw data, analyses files, and details of all packages used are available at https://github.com/aenneb/image_and_music_pleasure. This study was not preregistered.

Correspondence concerning this article should be addressed to Amy M. Belfi, Department of Psychological Science, 135 H-SS, 500 W. 14th St., Rolla, MO 65409, amybelfi@mst.edu

Abstract

Observers can make independent aesthetic judgements of at least two images presented briefly and simultaneously. However, it is unknown whether this is the case for two stimuli of different sensory modalities. Here, we investigated whether individuals can judge auditory and visual stimuli independently, and whether stimulus duration influences such judgments. Participants (*N*=120, across two experiments and a replication) saw images of paintings and heard excerpts of music, presented simultaneously for 2 s (Experiment 1) or 5 s (Experiment 2). After the stimuli were presented, participants rated how much pleasure they felt from the stimulus (music, image, or combined pleasure of both, depending on which was cued) on a 9-point scale. Finally, participants completed a baseline rating block where they rated each stimulus in isolation. We used the baseline ratings to predict ratings of audio-visual presentations. Across both experiments, the root-mean-square errors (RMSEs) obtained from leave-one-out-cross-validation analyses showed that people's ratings of music and images were unbiased by the simultaneously presented other stimulus, and ratings of both were best described as the arithmetic mean of the ratings from the individual presentations at the end of the experiment. This pattern of results replicates previous findings on simultaneously presented images, indicating that participants can ignore the pleasure of an irrelevant stimulus regardless of the sensory modality and duration of stimulus presentation.

Keywords: music; aesthetics; pleasure; reward

Word Count: 4,901

Public Significance Statement: This study suggests that people are able to make independent judgments of stimuli in their environment, for example, that a person can accurately judge how much they like a painting even if listening to a song they dislike.

Introduction

During everyday life, individuals regularly encounter multiple, simultaneous sources of sensory information. As an example, consider visiting a restaurant. While eating their meal, a diner is also exposed to sights (e.g., artwork, decor), sounds (e.g., conversations, music), and smells (e.g., other foods). Someone might walk away with a negative impression of a particular dish, but rather than being a "pure" judgment of the food itself, they may have been influenced by these other stimuli. In contrast to this naturalistic context in which people are exposed to multiple stimuli, most prior work has taken a reductionist approach to investigating aesthetic judgments. That is, participants are typically asked to make aesthetic judgments about items in a single sensory modality while eliminating all other sources of sensory input, such as judging an image of a single painting. Here, we sought to answer the question of whether individuals can make an independent aesthetic judgment of an item while simultaneously exposed to a stimulus in a different modality. To give another real-world example: If a person views their favorite painting while listening to a song they detest, do their negative feelings about the music influence their judgment of the painting?

Judgments of simultaneously presented images

Recent work sought to first address the question of whether observers can independently judge the pleasure of two stimuli presented in the *same* sensory modality. Researchers briefly presented two images side-by-side and observers were asked to rate the pleasure one of the images (Brielmann & Pelli, 2020). Based on predictions from prior literature, the authors tested multiple models for how observers might rate the pleasure of a single image in the presence of a second, distractor image. Additionally, the authors sought to investigate what happens when observers are asked to rate the combined pleasure of both images. They found that in in both cases, a “faithful” model best accounted for the data. That is, observers could both faithfully report the pleasure of a single image when accompanied by a second, irrelevant image, and also faithfully report the average of both images when asked to do so.

These findings (Berlmann & Pelli, 2020) stand in contrast to a line of work which demonstrates that people’s ratings of one item’s perceptual properties are biased toward the average of the set (e.g., Brady & Alvarez, 2011; Haberman & Whitney, 2009; Maule et al., 2014) They thus suggest that faithful reports of one’s *subjective evaluation* of an object might be different from reporting the object’s perceptual properties. However, it is not entirely clear whether pleasure judgments of multiple objects are unbiased under all circumstances. With longer presentation durations of 1.5 s and a homogeneous image set, for instance, pleasure ratings of one scene among three others are biased toward the mean pleasure of the presented scenes (Alwis & Haberman, 2020). In a follow-up study designed to test the limits in terms of the number of concurrent stimuli, Brielmann and Pelli (2021) repeated the same task but with four simultaneously presented images. The authors found that participants could *not* faithfully report the pleasure of a single image unless cued to which image they would be rating ahead of time, and that observers also did not faithfully report the average pleasure of four images (Brielmann & Pelli, 2021).

Thus, there seem to be clear limits to people’s ability to independently report the pleasure of multiple objects that they encounter simultaneously. So far, these limits have only been tested within the visual modality. However, another enticing question is whether people can encode and report the pleasure of two objects of different modalities that are presented at the same time. In the current study we sought to systematically investigate how pleasure judgments of simultaneously presented music and images are combined to form pleasure judgments of the image, the music, and the combination of both.

Judgments of simultaneously presented images and music

Prior literature suggests that observers may be influenced by a simultaneously presented auditory stimulus when asked to make judgments of a visual stimulus (for review, see Gerdes et al., 2014). For example, short excerpts of music have been shown to modulate judgments of the emotion depicted in facial expressions or other images (Logeswaran & Bhattacharya, 2009; Marin et al., 2012). Similarly, prior studies have investigated the influence of auditory stimuli on judgments of properties of visual stimuli: Listening to positively-valenced music is associated with perceiving a visual stimulus to be brighter (Bhattacharya & Lindsen, 2016).

Additionally, prior work indicates that listeners report stronger felt emotions when presented with emotionally congruent auditory and visual stimuli together than when each stimulus is presented in isolation (Baumgartner, Esslen, et al., 2006; Baumgartner, Lutz, et al., 2006; Pan et al., 2019). Additionally, Braun Janzen and colleagues (2022) demonstrated that museum visitors rated a Kandinsky painting as depicting more positive emotional valence if they listened to music they liked rather than disliked while viewing the painting. Similarly, Klein and colleagues (2021) reported that people like images with low complexity more in the presence of background music, independent of the music’s complexity. In sum, this work suggests that, in contrast to prior work presenting two visual stimuli, participants may be unable to faithfully judge a single stimulus in the presence of another when that stimulus is of a different sensory modality.

Influence of stimulus duration on aesthetic judgments

An additional component of the present work was to test what influence, if any, stimulus duration has on judgments of concurrently presented images and music. While previous research has indicated that observers can reliably report subjectively felt pleasure from an image in as little as 50 ms (Brielmann & Pelli, 2018; Forster et al., 2016; Schwabe et al., 2018), other prior work has indicated that listeners need at around 500-750 ms to accurately make aesthetic judgments of music (Belfi et al., 2018). Additionally, prior work found that observers cannot faithfully report the single pleasure of a visual stimulus in an array of four stimuli when presented very briefly (200 ms), but that they instead report overall diminished pleasure (Brielmann & Pelli, 2021). This result stands in contrast to similar work that presented multiple images for a longer duration (1.5 s) and found that observers are biased towards the mean pleasure of the entire array (Alwis & Haberman, 2020). When presented with multiple images for a longer period, observers likely had time to attend to each stimulus in a serial manner, which then subsequently resulted in averaging across all stimuli when making a judgment. Here, we expand on this prior work both by introducing two sensory modalities (visual, auditory) as well as two durations (2 s, 5 s).

The present study

The present study serves to clarify the nature of when and how sensory inputs from multiple sensory domains are combined in order to make evaluative judgments. Our current study therefore has theoretical implications for how we think about the way in which sensory information is combined. Brielmann and Pelli (2020, 2021) suggested that rating variance is a product of the rating process and that therefore, pleasures from two simultaneously presented images do not seem to be independently sampled. Here, we present stimuli of two different modalities, making it far more likely that pleasure is being sampled independently from the two sources. One goal of the present work is to provide an additional test for the hypothesis that, when combining pleasures, rating variability eschews during the rating process. The theoretical import of our study also reaches beyond sensory perception because the existence of separable (hedonic) information from simultaneously presented stimuli forms the basis for making value-based decisions. If it were true that people cannot report the pleasure of simultaneously presented auditory and visual objects, they would not be able to decide whether they want to continue engaging with either both, one, or none.

At a higher-order level, our work also seeks to inform several major theories of aesthetic judgments, particularly on the role played by extra-stimulus information. One of the major topics of debate in this field surrounds the relative contributions of stimulus features versus external factors to aesthetic judgments of sensory stimuli. Recent theories highlight the important role of context and extra-stimulus features and their influence on aesthetic or hedonic judgments of a stimulus (Brattico & Pearce, 2013; Leder & Pelowski, 2021; Pelowski et al., 2017). For example, prior work has found differences in aesthetic judgments based on the title of an artwork (Turpin et al., 2019), the performer of a musical piece (Belfi et al., 2021), and whether the composer of a musical piece is familiar or unknown (Fischinger et al., 2018). This work has indicated that, when asked to make an aesthetic judgment, observers tend to take a holistic approach such that extra-stimulus information does influence their ratings of a stimulus. While theories based on these findings suggest that contextual or external information should contribute to judgments of a single stimulus, here we sought to test whether this biasing by extra-stimulus information is compulsory. That is, when asked to rate a single stimulus in the presence of others (while ignoring the other ‘distractor’ stimuli), can observers faithfully do this? Or, is the contribution of extra-stimulus information obligatory? Thus, the findings from the present work will inform and refine such theories in terms of the role that extra-stimulus information plays in aesthetic judgments.

Therefore, in the present work, we sought to test 1) whether observers can independently judge the pleasure of a single stimulus (i.e., a single image or a single musical excerpt) when concurrently presented with a stimulus of the other sensory modality. Additionally, we sought to test whether observers can accurately average their pleasure responses to two stimuli when asked to rate the combined pleasure of the image and music. Here, participants viewed images of paintings and heard excerpts of instrumental music and were asked to rate the pleasure they felt either from a single stimulus or the combined pleasure of both stimuli. Additionally, we sought to investigate whether stimulus duration has an influence, and tested this question by repeating the experiment with short (2 s) and longer (5 s) stimulus durations.

Method

Participants

Participants were undergraduate students at Missouri University of Science and Technology who completed the experiments as part of their course credit. We were unable to conduct a power analysis to determine sample size because we do not employ traditional frequentist analyses and use specially designed models to fit our data. Since these models additionally differ in their complexity, i.e., number of free parameters, a simple frequentist test for differences between errors would be misleading, too. However, our current work is modeled very closely on our prior work (Brielmann & Pelli, 2020, 2021). We therefore set our target *N* at 30 participants, based on the sample sizes that have been used successfully in the past on a very similar task. The similar sample size additionally allows us to make direct comparisons between our and these previous findings. Note that we also employ our main analyses on the level of individual participants and that we can show that our results hold for the majority of individual participants, too. We additionally ensure the reliability of our results by replicating our experiment twice. To account for participant attrition due to potentially inattentive participants, we sought to recruit 40-50 participants per study.

Experiment 1

Of the 42 participants who completed Experiment 1, five were excluded from the analyses for failing the attention check. This left a total of 37 participants who completed the experiment (26 identified as male, 10 as female, and one as "other"). Their mean age was 19.8 years (*SD* = 2.1), they had an average of 13.66 (*SD*=1.3) years of education, 4.5 years (*SD*=3.42) of formal musical training, and 1.3 years (*SD*=1.71) of formal art training.

Experiment 2

Of the 48 participants who completed Experiment 2, nine were excluded from the analyses for failing the attention check. This left a total of 37 remaining participants, (27 identified as male and 12 as female). Their mean age was 19.4 years (*SD*=1.3), they had on average 13.7 years of education (*SD*=1.3), 3.7 years (*SD*=3.0) of formal musical training, and 2.0 years (*SD*=2.8) of formal art training. We repeated all analyses for both experiments with the removed subjects included, and the results remained unchanged. Therefore, below we report the results without these participants.

Stimuli

Stimuli consisted of 18 musical excerpts and 18 images of paintings. All stimuli were selected from our prior work (Belfi, 2019; Belfi et al., 2018, 2019). Musical stimuli consisted of instrumental musical excerpts selected from two genres: "classical" and "jazz." Classical excerpts were chosen from 19^{th} century small ensemble music of the Romantic era. Jazz music consisted of 1960s music with bop or postbop elements. Ensembles and instrumentation were all traditional jazz, consisting of piano, guitar, saxophone, brass, and so forth. No particularly well-known pieces were selected, in an effort to ensure that all pieces were unfamiliar. Musical pieces had previously been rated on their aesthetic appeal (Belfi et al., 2018). In order to span the range of aesthetic appeal in the present study, we selected the nine musical pieces that were the most highly rated (e.g., most liked) and the nine musical pieces that were the least highly rated (e.g., least liked).

Visual stimuli consisted of images of paintings that had been used in our prior work (Belfi et al., 2019). Images were selected from the Catalog of Art Museum Images Online database (http://www.oclc.org/camio). These images are high-quality photographs of paintings from a variety of cultural traditions (America, Asian, European) and time periods (15th century to the present). Commonly reproduced images and particularly well-known paintings were not included in order to minimize familiarity. As with musical stimuli, we selected the top nine most-liked and least-liked images based on previous ratings.

Procedure

All procedures were conducted in compliance with the American Psychological Association Ethical Principles and were approved by the Institutional Review Board of the University of Missouri. This study was conducted online using Gorilla Online Experiment Builder (http://www.gorilla.sc) to create and host the experiments described here (Anwyl-Irvine et al., 2020). Data was collected throughout the year 2021. Procedures generally followed those from a prior similar experiment using two images (Brielmann & Pelli, 2020) with some modifications for the present experiments. There were two blocks in the main task: a *precued* block and a *postcued* block. In the precued block, a cue was presented before the stimuli that indicated whether to rate the pleasure of the music, the image, or the combined pleasure of both stimuli. The cue was a word presented in the middle of the screen that said “music,” “image,” or “both.” In the postcued block, the same cues were presented but *after* the stimuli. The order of the precued and postcued blocks was counterbalanced across participants. At the end of these two blocks, participants completed a final baseline rating task. In this baseline task, only one stimulus appeared at a time and participants rated each stimulus in isolation. Baseline ratings were blocked, such that participants rated all images first and then all music (or vice versa; order of image and music baseline ratings were counterbalanced across participants).

First, participants were instructed on the task (see **Supplementary Materials** for the complete instructions). Next, they completed six practice trials, one for each possible cue (music, image, both), once precued and once postcued. Next, participants proceeded to the main experiment. Participants completed one precued block and one postcued block. In each block, each stimulus was presented once as *target* (i.e., the stimulus that is cued), once as *distractor* (i.e., the non-cued stimulus), and once as part of a pair whose combined pleasure was rated (total N=54 trials per block). At the beginning of each block, participants were told whether the cue would come before or after the stimuli. Finally, participants completed the baseline rating blocks. To rule out the possibility that baseline ratings at the end of the experiment were systematically corrupted by such sequence effects, we assessed whether the ratings of a participant or for a particular stimulus changed during the time course of the experiment (see **Supplementary Material** for details). Median correlations between trial number and ratings across participants or stimuli never exceeded an absolute value of 0.12 in any of the experiments, confirming that no overall systematic sequence effect could bias the relationship between ratings in the main experiment and baseline ratings.

Trial timelines for the different types of blocks are illustrated in **Figure 1.** Stimuli were presented for 2000 ms (Experiment 1) or 5000 ms (Experiment 2). Images were presented in the center of a white screen. After the stimuli, participants were instructed to "Rate how much pleasure you felt from the stimulus, where 1 is "no pleasure at all" and 9 is "very intense pleasure." Ratings were made by using a mouse click to select one of nine buttons labeled 1-9 on the screen. One attention check question was included, which was an audio stimulus that instructed participants to select a specific number on the rating scale.

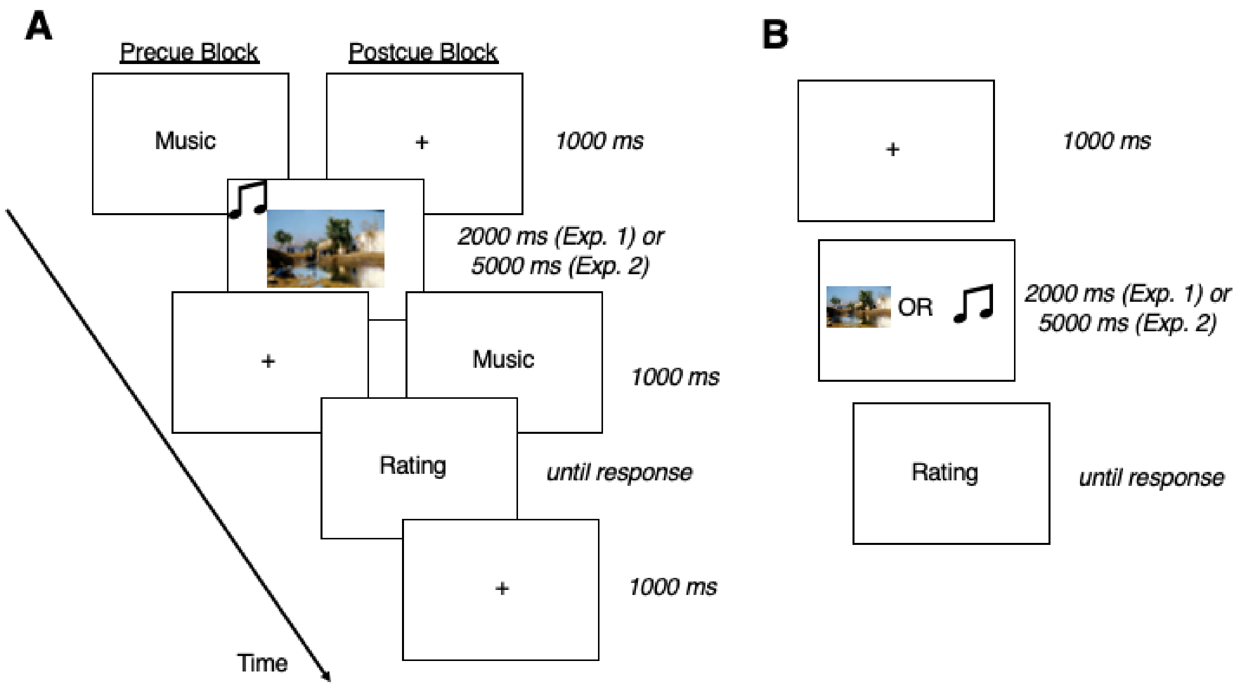

Figure 1. Timeline for one example trial for the main experiment (**A**) and the final baseline rating block (**B**).

At the end of the experiment, participants also completed two questionnaires to assess their overall pleasure responses to both music and visual images: The Barcelona Music Reward Questionnaire (BMRQ; Mas-Herrero et al., 2013) was used to assess musical reward, while the Aesthetic Responsiveness Assessment (AReA; Schlotz et al., 2020) was used to assess responsiveness to visual art. Finally, participants completed a brief demographics questionnaire and were asked about their age, gender, years of education, years of formal musical training, and years of formal visual arts training.

Analysis

We refer to trials in the main experiment where participants rated the pleasure of one target stimulus in the presence of a second distractor stimulus (*e.g.*, rating the music while an image was present) as *one-pleasure* trials. Trials in which participants rated the combined pleasure of both stimuli are referred to as *combined-pleasure* trials. Finally, the baseline trials at the end of the experiment are referred to as *single-pleasure* trials.

Our main analyses replicate the ones applied to data for two simultaneously presented images (Brielmann & Pelli, 2020). That is, we use leave-one-out-cross-validation (LOOCV) with the root-mean-squared error (RMSE) as our measure of goodness of fit to evaluate all models on an individual participant basis. As in the previous report, we model our data as a linear transformation of the weighted sum of both stimuli's single-pleasures:

$$\hat{P} = a + b(wP_1 + (1 - w)P_2) \quad (1)$$

where $\hat{P}$ is the estimated rated pleasure, w is the target weight, $0.5 \leq w \leq 1$, and a and b are constants. In one-pleasure trials, P_1 represents the target's and P_2 the distractor's single-pleasure; for combined-pleasure trials P_1 represents the image's single-pleasure and P_2 the music's single-pleasure.

In addition to the above models, we here also consider a modality-specific weighted sum of the stimuli's single-pleasures:

$$\hat{P} = wP_i + (1 - w)P_m \quad (2)$$

where P_i represents the image's single-pleasure and P_m the music's single-pleasure. Note that this equation is equivalent to Eq. 1 in combined-pleasure trials.

We fit four models to each participant's data. Three were based on Eq. 1: (1) the faithful model (w=1; a=0; b=1), (2) the compulsory averaging model (w=0.5; a=0; b=1), and (3) a linear model based on target assignment (a=0; b=1; where w is a free parameter). The fourth model (see Eq. 2) is also a linear model but assigns the weight w based on the stimulus modality (image vs. music).

For combined-pleasure trials, we considered three models that are variations of Eq. 1 assuming equal weighting of both stimulus pleasures (w = 0.5): (1) the faithful model with $a = 0$; $b = 1$; (2) the compressive model with $0 < b < 1$, $a > 0$; (3) the expansive model with $b > 1$; $a < 0$. In contrast, w is a free parameter for the fourth model we consider, the modality-specific linear model that weighs single-pleasures based on stimulus modality (Eq. 2).

Like previous studies (Brielmann & Pelli, 2020; Haberman et al., 2015) we report Cronbach's α for the average absolute error per participant as measure of reliability. Cronbach's alpha measures intercorrelations among items to gauge internal consistency. That is, for one-pleasure trials, we calculated the absolute deviation of each trial's pleasure rating from the single-pleasure rating (*i.e,* baseline rating) of the to-be-rated stimulus. For combined-pleasure trials we calculated absolute deviation of pleasure reports from the mean across single-pleasures of the two presented stimuli. We then averaged absolute errors per participant. Then, Cronbach's α values were used to compute an upper bound for the correlation between errors in the two kinds of trials, using the following equation (Eq. 3). In this equation, X and Y are two random variables (in our case, the errors in two kids of trials) and α is Cronbach's alpha (Nunnally, 1970):

$$r_{X,Y} \geq \sqrt{\alpha_X \alpha_Y} \qquad (3)$$

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in this study. Raw data, analyses files, and details of all packages used are available at https://github.com/aenneb/image_and_music_pleasure. Materials are copyrighted and therefore available upon request. We analyzed the data with Python 3.8.1 running in Spyder 4.1.4. This study's design and its analysis were not pre-registered.

Results

Experiment 1: Short stimulus presentation

Reliability and correlations between different trial types

Cronbach's alpha varied considerably between participants but was, on average, high across trial types (see **Figure 2A**). In contrast to what has been reported for two images, reliability tended to be higher for combined-pleasure trials in our mixed-modality experiment (precued $\alpha = 0.81$, 95% CI [0.72, 0.89]; postcued $\alpha = 0.84$, [0.76, 0.91]) than for one-pleasure trials (precued $\alpha = 0.45$ [0.17, 0.68]; postcued $\alpha = 0.66$[0.49, 0.80]). The non-overlapping confidence intervals (CIs) for precued trials indicate that this difference is meaningful at least in this case. The fact that reliability of combined ratings were higher than for one-target ratings is expected under the assumption that these ratings represent an averaged report of two individually sampled pleasure values. Note that this result stands in contrast to equal (if not higher) reliability of single-pleasure reports for two images (Brielmann & Pelli, 2020).

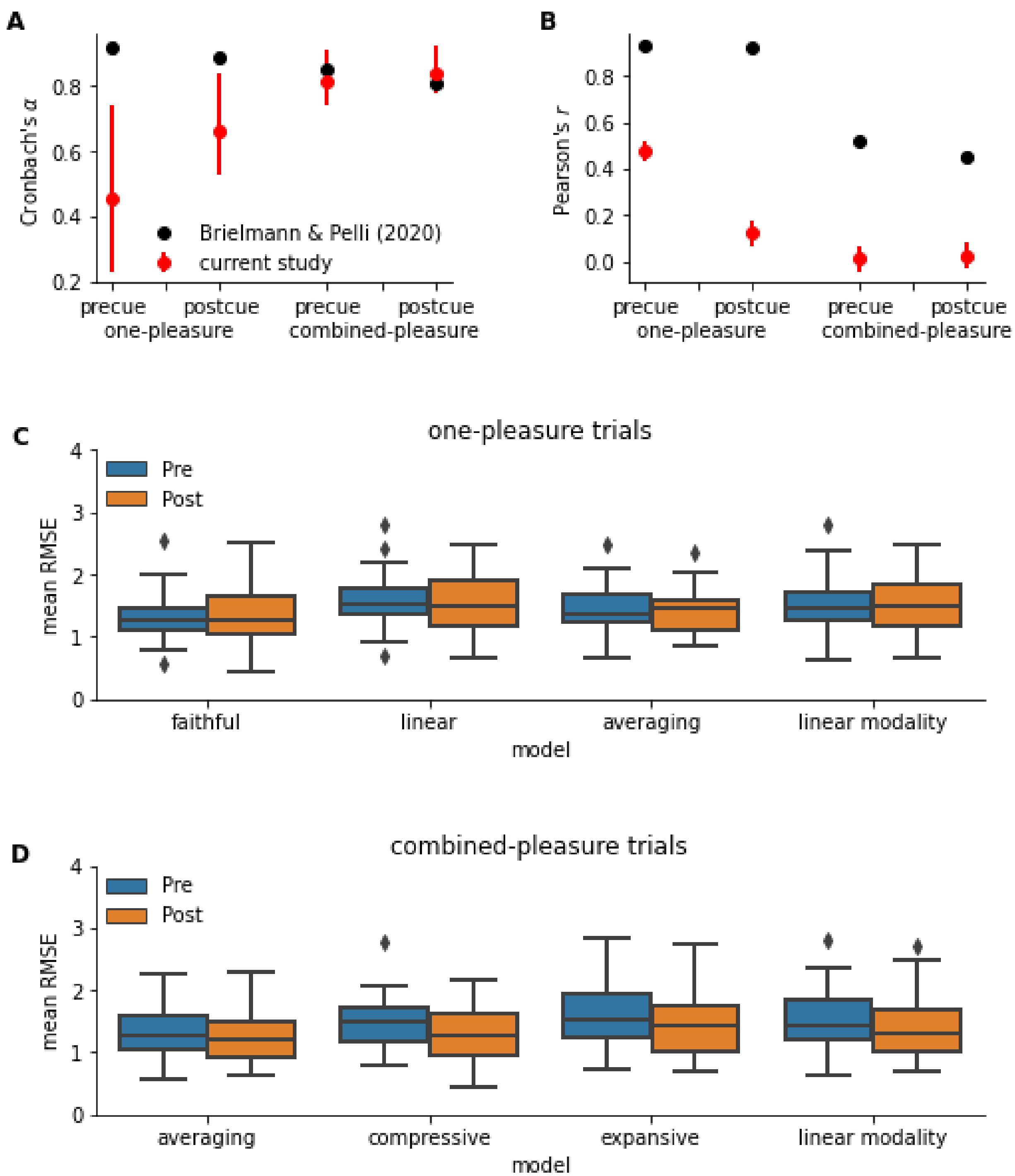

Figure 2. Reliability, correlations, and model fitting results for experiment 1. A) Cronbach's α for rating errors as a measure of reliability per trial type. b) Pearson's *r* as measure of correlation between rating errors between trial types. A-B) Red dots represent data of the current study, black ones of a previous study on two simultaneously presented images (Brielmann & Pelli, 2020). Error bars indicate 95% confidence intervals. C-D) Boxplots of root-mean-squared errors (RMSEs) for each model for precued trials (left; blue) and postcued trials (right; orange). Boxplot properties are the default setting of the python package *seaborn*. C) RMSEs for one-pleasure trials. D) RMSEs for combined-pleasure trials.

Inserting the above Cronbach alpha values into Equation 3, the maximal expected correlations are 0.55 between pre- and postcued one-pleasure trials, 0.83 between pre- and postcued combined-pleasure trials, 0.61 between one- and combined precued trials, and 0.75 between one- and combined-pleasure postcued trials. As illustrated in **Figure 2B**, correlations between trial-types were well below the maximally achievable correlations and considerably lower in our study than previously reported for two images. Nonetheless, the relative differences between trial-types were similar in both studies. Error correlations were higher between trials with the same instructions (rate-one r=0.47, 95% CI[0.43, 0.51]; rate both r=0.12, [0.07, 0.18]) than between trials with the same cue timing (pre-cued r=0.01, [-0.04, 0.07]; post-cued r=0.03, [-0.03, 0.08]).

People can report the pleasure of simultaneously presented images and music independently

Figure 2C compares the outcomes of all models for single-pleasure trials. On average, the faithful model fit the data in pre- (mean RMSE = 1.31) and postcued trials (1.32) best and it was the nominally best-fitting model according to RMSE for the majority of participants in pre- (26/37) and postcued trials (24/37), too. Thus, people report the pleasure of one image or musical piece in an unbiased fashion even when accompanied by music or an image respectively.

People can report the average pleasure across an image and music

Figure 2D compares the outcomes of LOOCV for combined-pleasure trials for all models. On average, the faithful averaging model fit the data in pre- (mean RMSE = 1.22) and postcued trials (1.30) best and it was the best-fitting model according to RMSE for the majority of participants in pre- (26/37) and postcued trials (21/37), too. No other model fit a similar number of participants best (see **Supplementary Material** for detailed counts). Thus, people's judgments of the combined pleasure they experience based on an image accompanied by music are best captured as the arithmetic average of the music's and the image's single-pleasure ratings.

We also investigated whether there were systematic relationships between the goodness of model fits and/or parameter estimates and participants' BMRQ and AReA scores. By and large, there were no meaningful and consistent correlations. We report the detailed analyses and results in the **Supplementary Material**.

Experiment 2: Long stimulus presentation

The analyses for Experiment 2 were identical to those for Experiment 1. As illustrated in **Figure 3**, the increased stimulus duration did not change the results. Cronbach's alpha was again slightly lower in one-pleasure trials (precued $\alpha = 0.68$, 95% CI [0.51, 0.81]; postcued $\alpha = 0.75$, [0.62, 0.85]) than in combined-pleasure trials (precued $\alpha = 0.81$, [0.71, 0.89]; postcued $\alpha = 0.83$, [0.74, 0.90]). However, the overlapping confidence intervals suggest that the differences in this experiment were less robust. Error correlations were again higher between trials with the same instructions (rate-one r=0.43, [0.39, 0.47]; rate both r=0.12, [0.07, 0.17]) than between trials with the same cue timing (pre-cued r=-0.07, [-0.12, -0.02]; post-cued r=0.03, [-0.03, 0.08]).

The faithful model achieved the on average lowest error in single-pleasure trials (precued RMSE = 1.26; postcued RMSE = 1.32) and was the best-fitting model for most participants (29/39 precued; 27/39 postcued). Also, the faithful averaging model had the lowest average RMSE across participants in combined-pleasure trials (both RMSEs = 1.37) as well as being the best-fitting model for a majority of individual participants (24/39 precued; 22/39 postcued).

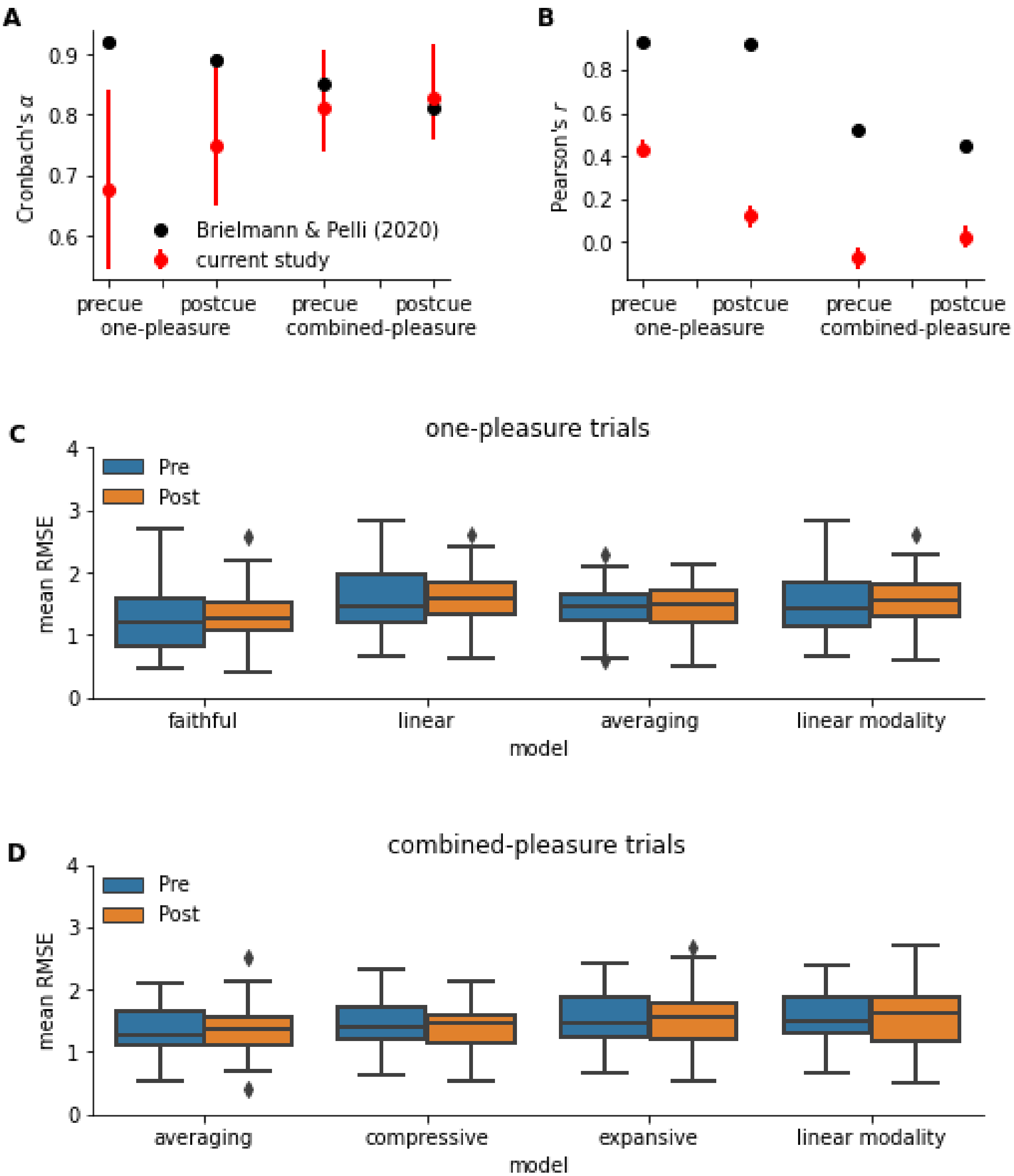

Figure 3. Reliability, correlations, and model fitting results for experiment 2. A) Cronbach's α for rating errors as a measure of reliability per trial type. b) Pearson's *r* as measure of correlation between rating errors between trial types. A-B) Red dots represent data of the current study, black ones of a previous study on two simultaneously presented images (Brielmann & Pelli, 2020). Error bars indicate 95% confidence intervals. C-D) Boxplots of root-mean-squared errors (RMSEs) for each model for precued trials (left; blue) and postcued trials (right; orange). Boxplot properties are the default setting of the python package *seaborn*. C) RMSEs for one-pleasure trials. D) RMSEs for combined-pleasure trials.

Discussion

In the present work, we sought to investigate whether individuals can report the pleasure of a single item (i.e., a musical excerpt or image of a painting) in the presence of a stimulus in the other modality – That is, can people accurately judge the pleasure of an image while listening to music, and vice versa? Our results indicate that people *can* do this: Participants reported the pleasure of an image or a piece of music faithfully, even in the presence of the other. Additionally, we sought to test whether people can accurately average the pleasure of two stimuli when asked to do so. Again, our results indicated that people can faithfully report the average: When prompted to rate their combined pleasure, people's reports are best described as the arithmetic mean of the image's and the music's individually rated pleasures. This pattern of results was evident whether people knew what to rate before stimulus onset (i.e., the precued trials) or after stimulus offset (i.e., the postcued trials) as well as with shorter (2 s) and longer (5 s) exposure durations. To further illustrate the robustness of this effect, we also replicated these main results (using the shorter, 2 s stimulus presentation) in an online sample of participants (see Experiment 3 in **Supplementary Materials**).

These findings also replicate those previously reported for studies with two simultaneously presented images (Brielmann & Pelli, 2020). This suggests that people can faithfully encode and report the pleasures of at least two objects, independent of correspondence in stimulus modality. While our main results replicated those of image-image pairings, we did find some differences between the current and the previous image-only studies. For one, rating reliability was higher for combined-pleasure ratings (as compared to one-pleasure ratings) in the studies reported here, as one would expect when pleasures are averaged across two independent observations, whereas it was no different for one- versus combined-pleasure ratings of two images. This suggests that late-stage response noise arises when averaging pleasures of the same modality but not when combining pleasures of two different modalities. In addition, correlations between average errors in different trial types were considerably lower in our study than they could have been based on their reliability. This, too, stands in contrast to previous findings with two images where correlations were at or close to ceiling. Hence, rating behavior varied more as a function of cue timing and requested report (one versus combined pleasure) when people experienced an image and music rather than two images. This suggests that participants used different strategies (implicitly or explicitly) in the different trial types. Nonetheless, our modeling results show that all of these presumably different strategies resulted in faithful pleasure reports of the target stimulus and the average across stimuli in one- and combined-pleasure trials respectively.

At first glance, it might seem that this work stands in some contrast to prior work investigating how the presence of music influences judgments of visual stimuli. That is, prior work found that simultaneously presented music does have an influence on judgments of images. For example, hearing a “happy” song while viewing a “happy” face was associated with increased judgments of positive valence, as compared to hearing the music alone (Pan et al., 2019) or seeing the face alone (Baumgartner, Lutz, et al., 2006). However, our study was distinct from these prior works in the type of response participants made. Here, we were interested in the amount of pleasure felt in response to the stimuli, rather than judgments of the emotional *content* of the stimuli. This contrast between studies of pleasure ratings versus ratings of stimulus features is similar to that found in Brielmann and Pelli’s findings (2020). While they found that observers could reliably report the *pleasure* of a single item in the presence of others, this stands in contrast to prior work demonstrating that ratings of a single item’s *perceptual properties* are biased toward the average of the set, rather than faithfully reported (e.g., Brady & Alvarez, 2011; Haberman & Whitney, 2009; Maule et al., 2014). Another difference is that here, we randomly combined pairs of most-pleasing and least-pleasing music and images (as determined by ratings from prior work), whereas prior work chose images and music to depict certain emotions or stimulus characteristics (e.g., happiness vs. sadness, or complex vs. simple). Thus, it seems to be that rating features of the stimulus (whether perceptual or emotional) may be more at risk to bias from other simultaneously presented stimuli than rating one's felt pleasure in response to a stimulus.

An additional difference between the present study and prior, similar work, is that previous studies tend to focus on the influence of "background" music (Braun Janzen et al., 2022; Klein et al., 2021). These prior studies presented a single extended musical excerpt, for minutes or longer, while the visual stimuli changed on screen (Baumgartner, Esslen, et al., 2006; Baumgartner, Lutz, et al., 2006). While we found no difference between 2 s and 5 s stimulus presentation, such very extended presentation durations may lead to one stimulus influencing judgments of the other. Therefore, it could be the case that one's rating of an image would be influenced by a concurrent musical stimulus if presented for an extended period of time.

One important caveat of the present work is that it can only be generalized to certain populations and certain types of stimuli. To generalize beyond an undergraduate population, we replicated Experiment 1 with participants recruited online via Prolific (see **Supplementary Material**). Prior work has indicated that online samples tend to be more diverse than undergraduate student samples, and has additionally shown that Prolific samples tend to be more diverse than samples recruited on other online platforms such as Amazon's Mechanical Turk (Peer et al., 2017). However, we limited this online sample to US-based participants, to match our US-based undergraduates. Therefore, we cannot generalize our results beyond a US-based population. We here aim to make claims about this population's ability to report pleasure from simultaneously presented visual and auditory stimuli and therefore did not recruit additional expert populations, e.g., musicians or art historians, and leave it up to future studies to examine whether our results will generalize to such special populations. Additionally, it is important to note that for our musical stimuli, we used classical and jazz musical excerpts that come from the Western musical tradition. This is a limitation that does not allow generalizability beyond Western listeners of Western music (Baker et al., 2020).

To conclude, our results show that individuals can give unbiased reports of their felt pleasure from an image or a piece of music, even when in the presence of another stimulus. Additionally, they are also able to faithfully report the average pleasure of both stimuli when asked to do so. These results replicate findings with two image pairings and extend this work into judgments of multisensory stimuli. Future research could further extend these findings into other sensory domains (e.g., touch, smell) and continue to probe the question of stimulus duration in order to test the limits of one's ability to independently track the pleasure of simultaneous stimuli.

Open Practices Statement

Raw data, analyses files, and details of all packages used are available at https://github.com/aenneb/image_and_music_pleasure. None of the experiments reported here were preregistered.

References

Alwis, Y., & Haberman, J. M. (2020). Emotional judgments of scenes are influenced by unintentional averaging. *Cognitive Research: Principles and Implications*, *5*(1). https://doi.org/10.1186/s41235-020-00228-3

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baker, D. J., Belfi, A., Creel, S., Grahn, J., Hannon, E., Loui, P., Margulis, E. H., Schachner, A., Schutz, M., Shanahan, D., & Vuvan, D. T. (2020). Embracing anti-racist practices in the music perception and cognition community. *Music Perception*, *38*(2), 103–105. https://doi.org/10.1525/mp.2020.38.2.103

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychophysiology*, *60*(1), 34–43. https://doi.org/10.1016/j.ijpsycho.2005.04.007

Baumgartner, T., Lutz, K., Schmidt, C. F., & Jäncke, L. (2006). The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research*, *1075*(1), 151–164.

Belfi, A. M. (2019). Emotional valence and vividness of imagery predict aesthetic appeal in music. *Psychomusicology: Music, Mind, and Brain*, *29*, 128–135.

Belfi, A. M., Kasdan, A., Rowland, J., Vessel, E. A., Starr, G. G., & Poeppel, D. (2018). Rapid timing of musical aesthetic judgments. *Journal of Experimental Psychology: General*, *147*(10), 1531–1543. https://doi.org/http://dx.doi.org/10.1037/xge0000474

Belfi, A. M., Samson, D. W., Crane, J., & Schmidt, N. L. (2021). Aesthetic judgments of live and recorded music: Effects of congruence between musical artist and piece. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.618025

Belfi, A. M., Vessel, E. A., Brielmann, A., Isik, A. I., Chatterjee, A., Leder, H., Pelli, D. G., & Starr, G. G. (2019). Dynamics of aesthetic experience are reflected in the default-mode network. *NeuroImage*, *188*. https://doi.org/10.1016/j.neuroimage.2018.12.017

Bhattacharya, J., & Lindsen, J. P. (2016). Music for a brighter world: Brightness judgment bias by musical emotion. *PLoS ONE*, *11*(2), 1–11. https://doi.org/10.1371/journal.pone.0148959

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brattico, E., & Pearce, M. (2013). The neuroaesthetics of music. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(1), 48–61. https://doi.org/10.1037/a0031624

Braun Janzen, T., de Oliveira, B., Ventorim Ferreira, G., Sato, J. R., Feitosa-Santana, C., & Vanzella, P. (2022). The effect of background music on the aesthetic experience of a visual artwork in a naturalistic environment. *Psychology of Music*, 1–17. https://doi.org/10.1177/03057356221079866

Brielmann, A. A., & Pelli, D. G. (2018). Aesthetics. *Current Biology*, *28*(16), R859–R863. https://doi.org/10.1016/j.cub.2018.06.004

Brielmann, A. A., & Pelli, D. G. (2020). Tracking two pleasures. *Psychonomic Bulletin and Review*, *27*(2), 330–340. https://doi.org/10.3758/s13423-019-01695-6

Brielmann, A. A., & Pelli, D. G. (2021). The pleasure of multiple images. *Attention, Perception, and Psychophysics*, *83*(3), 1179–1188. https://doi.org/10.3758/s13414-020-02175-z

Fischinger, T., Kaufmann, M., & Schlotz, W. (2018). If it's Mozart, it must be good? The influence of textual information and age on musical appreciation. *Psychology of Music*, *48*, 579–597. https://doi.org/10.1177/0305735618812216

Forster, M., Leder, H., & Ansorge, U. (2016). Exploring the subjective feeling of fluency. *Experimental Psychology*, *63*.

Gerdes, A. B. M., Wieser, M. J., & Alpers, G. W. (2014). Emotional pictures and sounds: A review of multimodal interactions of emotion cues in multiple domains. *Frontiers in Psychology*, *5*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2014.01351

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., & Whitney, D. (2009). Seeing the Mean: Ensemble Coding for Sets of Faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734. https://doi.org/10.1037/a0013899

Klein, K., Melnyk, V., & Völckner, F. (2021). Effects of background music on evaluations of visual images. *Psychology and Marketing*, *38*(12), 2240–2246. https://doi.org/10.1002/mar.21588

Leder, H., & Pelowski, M. (2021). Empirical Aesthetics: Context, extra information, and framing. In M. Nadal & O. Vartanian (Eds.), *The Oxford Handbook of Empirical Aesthetics*. https://doi.org/10.1093/oxfordhb/9780198824350.013.43

Logeswaran, N., & Bhattacharya, J. (2009). Crossmodal transfer of emotion by music. *Neuroscience Letters*, *455*(2), 129–133. https://doi.org/10.1016/j.neulet.2009.03.044

Marin, M. M., Gingras, B., & Bhattacharya, J. (2012). Crossmodal transfer of arousal, but not pleasantness, from the musical to the visual domain. *Emotion*, *12*(3), 618–631.

https://doi.org/10.1037/a0025020

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception*, *31*, 118–138.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America A*, *31*(4), A93. https://doi.org/10.1364/josaa.31.000a93

Nunnally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill.

Pan, F., Zhang, L., Ou, Y., & Zhang, X. (2019). The audio-visual integration effect on music emotion: Behavioral and physiological evidence. *PLoS ONE*, *14*(5), 1–21. https://doi.org/10.1371/journal.pone.0217040

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me… delight my eyes and brain: The Vienna Integrated Model of top–down and bottom–up processes in Art Perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. *Physics of Life Reviews*, 80–125. https://doi.org/10.1016/j.plrev.2017.02.003

Schlotz, W., Wallot, S., Omigie, D., Masucci, M. D., Hoelzmann, S. C., Vessel, E. A., Wallot, S., & Omigie, D. (2020). The Aesthetic Responsiveness Assessment (AReA): A Screening Tool to Assess Individual Differences in Responsiveness to Art in English and German The Aesthetic Responsiveness Assessment (AReA): A Scr. *Psychology of Aesthetics , Creativity , and the Arts*.

Schwabe, K., Menzel, C., Mullin, C., Wagemans, J., & Redies, C. (2018). Gist Perception of Image Composition in Abstract Artworks. *I-Perception*, *9*(3), 204166951878079. https://doi.org/10.1177/2041669518780797

Turpin, M. H., Walker, A. C., Kara-Yakoubian, M., Gabert, N. N., Fugelsang, J. A., & Stolz, J. A. (2019). Bullshit makes the art grow profounder. *Judgment and Decision Making*, *14*(6), 658–670. https://doi.org/10.2139/ssrn.3410674

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychophysiology*, *60*(1), 34–43. https://doi.org/10.1016/j.ijpsycho.2005.04.007

Baumgartner, T., Lutz, K., Schmidt, C. F., & Jäncke, L. (2006). The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research*, *1075*(1), 151–164.

Belfi, A. M. (2019). Emotional valence and vividness of imagery predict aesthetic appeal in music. *Psychomusicology: Music, Mind, and Brain*, *29*, 128–135.

Belfi, A. M., Kasdan, A., Rowland, J., Vessel, E. A., Starr, G. G., & Poeppel, D. (2018). Rapid timing of musical aesthetic judgments. *Journal of Experimental Psychology: General*, *147*(10), 1531–1543. https://doi.org/http://dx.doi.org/10.1037/xge0000474

Belfi, A. M., Vessel, E. A., Brielmann, A., Isik, A. I., Chatterjee, A., Leder, H., Pelli, D. G., & Starr, G. G. (2019). Dynamics of aesthetic experience are reflected in the default-mode network. *NeuroImage*, *188*. https://doi.org/10.1016/j.neuroimage.2018.12.017

Bhattacharya, J., & Lindsen, J. P. (2016). Music for a brighter world: Brightness judgment bias by musical emotion. *PLoS ONE*, *11*(2), 1–11. https://doi.org/10.1371/journal.pone.0148959

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Braun Janzen, T., de Oliveira, B., Ventorim Ferreira, G., Sato, J. R., Feitosa-Santana, C., & Vanzella, P. (2022). The effect of background music on the aesthetic experience of a visual artwork in a naturalistic environment. *Psychology of Music*, 1–17. https://doi.org/10.1177/03057356221079866

Brielmann, A. A., & Pelli, D. G. (2018). Aesthetics. *Current Biology*, *28*(16), R859–R863. https://doi.org/10.1016/j.cub.2018.06.004

Brielmann, A. A., & Pelli, D. G. (2020). Tracking two pleasures. *Psychonomic Bulletin and Review*, *27*(2), 330–340. https://doi.org/10.3758/s13423-019-01695-6

Brielmann, A. A., & Pelli, D. G. (2021). The pleasure of multiple images. *Attention, Perception, and Psychophysics*, *83*(3), 1179–1188. https://doi.org/10.3758/s13414-020-02175-z

Forster, M., Leder, H., & Ansorge, U. (2016). Exploring the subjective feeling of fluency. *Experimental Psychology*, *63*.

Gerdes, A. B. M., Wieser, M. J., & Alpers, G. W. (2014). Emotional pictures and sounds: A review of multimodal interactions of emotion cues in multiple domains. *Frontiers in Psychology*, *5*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2014.01351

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., & Whitney, D. (2009). Seeing the Mean: Ensemble Coding for Sets of Faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734. https://doi.org/10.1037/a0013899

Klein, K., Melnyk, V., & Völckner, F. (2021). Effects of background music on evaluations of visual images. *Psychology and Marketing*, *38*(12), 2240–2246. https://doi.org/10.1002/mar.21588

Logeswaran, N., & Bhattacharya, J. (2009). Crossmodal transfer of emotion by music. *Neuroscience Letters*, *455*(2), 129–133. https://doi.org/10.1016/j.neulet.2009.03.044

Marin, M. M., Gingras, B., & Bhattacharya, J. (2012). Crossmodal transfer of arousal, but not pleasantness, from the musical to the visual domain. *Emotion*, *12*(3), 618–631. https://doi.org/10.1037/a0025020

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception*, *31*, 118–138.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America A*, *31*(4), A93. https://doi.org/10.1364/josaa.31.000a93

Nunnally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill.

Pan, F., Zhang, L., Ou, Y., & Zhang, X. (2019). The audio-visual integration effect on music emotion: Behavioral and physiological evidence. *PLoS ONE*, *14*(5), 1–21. https://doi.org/10.1371/journal.pone.0217040

Schlotz, W., Wallot, S., Omigie, D., Masucci, M. D., Hoelzmann, S. C., Vessel, E. A., Wallot, S., & Omigie, D. (2020). The Aesthetic Responsiveness Assessment (AReA): A Screening Tool to Assess Individual Differences in Responsiveness to Art in English and German The

Aesthetic Responsiveness Assessment (AReA): A Scr. *Psychology of Aesthetics , Creativity , and the Arts*.

Schwabe, K., Menzel, C., Mullin, C., Wagemans, J., & Redies, C. (2018). Gist Perception of Image Composition in Abstract Artworks. *I-Perception*, *9*(3), 204166951878079. https://doi.org/10.1177/2041669518780797

Learning Experience Reflection for 2023 Undergraduate Research Conference
Jessica C. Frame – UGRC 2023

Taking part in this research project has led me to get a strong foundational understanding of how research is conducted in psychology and in neuroscience. The pipeline for this field of research is relatively similar to many others. Research starts with a general question the investigator wishes to focus on. This typically can get kickstarted from knowledge of previous research and the want to expand into new horizons from that baseline. From here, a general study design is outlined and from that, hypotheses are formed. These hypotheses lead to more fleshed out study that is put through rounds of piloting to gain feedback on any problems that might have occurred during the design phase. After piloting, studies are released to a population that was decided on to be the focus in the study during the planning stage. Data is then collected until the investigator believes the amount of data is sufficient to be subjected to data analyses. From here a wide range of statistical tools are at the investigator's disposal to use to best analyze and pull conclusions from the data. From here, an investigator can either refute or fail to refute the hypotheses previously made. Generally, as seen above, the pipeline of research in psychology is rather standard but there is one thing that has to be done typically not seen in research that doesn't deal with human participants. Before a study in psychology that involves human participants-research has to be looked over and screened by an IRB and ethics committee to make sure the research doesn't exploit or affect participants in an unnecessarily negative way.

Through writing a manuscript, I have found a new understanding for knowing what sources you are using and where they're coming from. Having really good and thorough sources can really make or break the argument you're trying to set for your research. Informational sources are the backbone of your introduction that segways into your entire paper. It's imperative to set an notion that your questions and hypotheses have come to fruition from previous ideas that have flourished and have reasonable statistical and logical background to show that you've taken time to have an educated foundation for not only yourself as an investigator but for your research and those who see it.

Through this research I've conducted, I've gained a much better understanding on the fundamentals of experimental design. The biggest thing I've taken away is to truly understand what exactly you are measuring. This allows you to understand what type of statistical measures you should conduct to get the most out of your data and interpret it correctly in relation to what is being measured. Bouncing off of this, knowing thoroughly what you're studying allows you to back-track and remove any plausible confounds from the study or at least acknowledge its possible effects on your results.

From this research, I've learned many useful methods to understand and interpret my results. You first start off with your raw data and hypotheses, based on what you want to measure-you're then able to pick what statistical analyses might be best for the study you've conducted. Every corner of research has thresholds for what is considered significant data. If the data is significant, we can then say we fail to reject our hypothesis and can go into discussion and conclusions as to what this means. The same can be said for if we reject the hypothesis but this can lead to returning to the drawing board to a new experiment that may better encompass what the research being down is about if deemed necessary. I've learned to do analyses through statistical means and what they imply for the study that has been done.