Tendency Towards the Average? The Aesthetic Evaluation of a Quantitatively Average Music Performance: A Successful Replication of Repp's (1997) Study

Anna Wolf
Universität Hamburg, Hamburg, Germany

Reinhard Kopiez
Hanover University of Music, Drama and Media, Hanover, Germany

Friedrich Platz
University of Music and Performing Arts, Stuttgart, Germany

Hsin-Rui Lin & Hanna Mütze
Hanover University of Music, Drama and Media, Hanover, Germany

The present study has investigated the minimal-distance hypothesis in music (Langlois & Roggman, 1990; Repp, 1997) by replicating Repp's original study (1997) on the aesthetic quality of an averaged performance—compared to individual interpretations—of Robert Schumann's Träumerei (Op. 15, No. 7). Participants ($N = 205$) came from Germany and Taiwan and made up a convenience sample representing different degrees of musical sophistication. We used a 2 × 4 mixed methods design that compared the country of data collection (between factor) and the four selected interpretations (within factor). The dependent variable was a unidimensional construct describing the musical quality, which was developed with an exploratory factor analysis followed by a probabilistic item analysis. It was found that the evaluation of Taiwanese and German participants did not differ, but the ratings for the various interpretations successfully replicated Repp's results: The average performance was rated better than the individual performances, and the lowest rated performance from the original study was rated lowest in this replication as well (large effect size). The confirmation of this central effect in music perception research might be an incentive for further replication studies in music psychology.

Received October 12, 2017, accepted March 21, 2018.

Key words: empirical aesthetics, cross-cultural, performance, music, evaluation

When Leonardo da Vinci (1452-1519) was commissioned by the Duke of Milan in 1482 to manufacture an equestrian monument, he started with a series of systematic, anatomical studies of horses to reveal the secret of beauty in artworks. His work was guided by the idea that the ideal proportions of a monument are represented by the average proportions of individual horses, a concept that he explored in numerous sketches (Williams, 1966). However, the realization of the "Gran Cavallo" monument never got past the stage of a clay model and had to wait 500 years until it was finished in 1999. In the late 19th century, the English anthropologist Francis Galton (1879) first tested the assumption that an average portrait (a so-called composite)—resulting from the superimposition of single photographic portraits of several persons—might reveal the typical characteristics of a defined group of individuals. His intention was to produce a picture of men who are likely to become criminals, prefer a vegetarian diet, or are tuberculosis patients (for a discussion, see Langlois & Roggman, 1990). Surprisingly, Galton observed that the composite pictures looked much better than their components, which numbered up to eight. Research in the 20th century on physical attractiveness used Galton's approach and showed that computed average faces, male and female, were rated as highly attractive and more positively when compared to most of the individual portraits. As Langlois and Roggman (1990) found, attractiveness ratings increased linearly with the number of underlying faces (up to 32 in their experiments). This seems to be a stable effect across age groups and cultures. The preference for the average face version is usually explained by its prototypicality in terms of the smallest average distance from an individual's aesthetic ideals. This is called the *minimal-distance hypothesis* (MDH; Repp, 1997).

Music Perception, volume 36, issue 1, pp. 98–108, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2018.36.1.98

In the domain of music, only very few experiments have been conducted to test the validity of the MDH: In a rating experiment by Kopiez, Langner, and Steinhagen (1999), German and Ghanaian participants evaluated the quality of six short drum-pattern performances. For most of the patterns, those versions that were average in terms of timing and dynamics were evaluated best in both countries. The second (and to the best of our knowledge earliest) study on the aesthetic attractiveness of an average musical interpretation was conducted by Repp (1997). In this study, 10 piano students performed Robert Schumann's "Träumerei," Op. 15, No. 7 three times on a MIDI piano; MIDI files were edited (correction of wrong notes, synchronization of nominally simultaneous notes), and note onsets and MIDI velocities of the three versions were averaged, resulting in a single average performance per player. Note offsets were inserted so that articulation was legato throughout, and a uniform pedaling pattern was superimposed. Finally, based on the MIDI data of the 10 averaged individual performances, a grand average performance (AP) was created, and MIDI data were converted back into sound (Roland RD-250s digital piano). This means that performances only differed in timing (expressive timing and basic tempo) and dynamics. Participants (12 piano students) listened to randomized orders of the individual performances and the average performance and then gave an overall rating on an 11-point scale. The items used for evaluation were as follows: tempo (*much too slow–much too fast*), dynamics (*much too weak–much too strong*), expression (*very inexpressive–very exaggerated*), and individuality (*conventional–very unusual*) on a 5-point rating scale. Furthermore, an overall rating on an 11-point rating scale was given. Consistency of ratings in the first round (so-called "semifinals") was controlled for by a second evaluation (so-called "finals"). The main results only relied on the 11-point rating scale, which was used efficiently by the participants, as can be seen in Figure 1. The AP version was rated second highest and was only outperformed by the individual version P10. Version P8 received the lowest ranking. Unfortunately, no statistical analysis for between-version differences was given (e.g., paired comparisons); thus, findings can only be interpreted by the reader on the basis of ocular inspection. Moreover, the presumably interesting results of the four more specific 5-point scales were not reported at all. In a follow-up study (Experiment 2 with much shorter sections of a Chopin Etude of only 22 s duration), Repp (1997) could show that the AP version was ranked highest only if performances from outstanding pianists were considered in the averaging. The author concluded that "the results provide tentative support for the hypothesis that an average performance can sound more appealing than the majority of the performances that go into the average. The effect was more striking for the set of expert performances in Experiment 2 than for the student performances in both experiments" (p. 439).

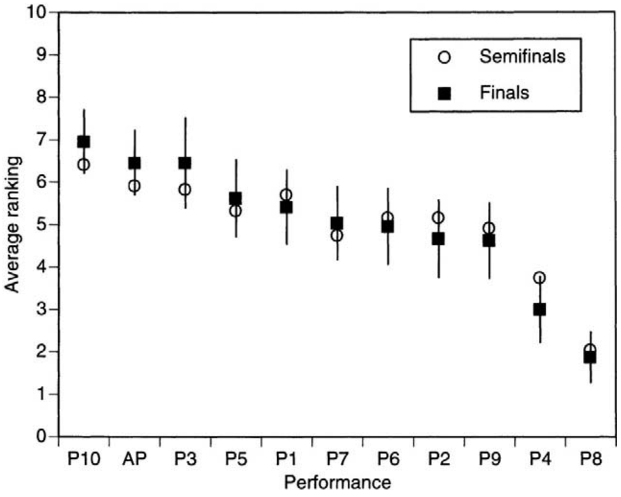

FIGURE 1. Average semifinal and final rankings of the 10 individual performances and the performance (AP), with standard error bars for the final ranking (from Repp, 1997; with kind permission of University of California Press).

What are the reasons for a replication of these findings? In our view, there are three main reasons for the replication of Repp's (1997) study: First, we think the study is of outstanding relevance for the field of empirical aesthetics and, thus, results should be verified twenty years after the initial publication. However, in our replication, the original study design was adopted and extended to the current state of research. Second, against the background of the "replication crisis" in psychology, we have to presume that only a daunting estimate of 39% studies can be successfully replicated, and that overall, half the effect size as in the original study can be found again (Open Science Collaboration, 2015). Both Ioannidis (2014) and Cumming (2014) reported extensive lists with mostly viable suggestions on how to improve current research practice to reduce this problem of low replicability. These include the sharing of materials, the usage of effect sizes and confidence intervals, and the choice of appropriate statistical methods—especially contrary to the random application of null hypothesis significant testing (NHST). On a more positive note, Klein et al. (2014) managed to replicate 10 of 13 classic psychological effects, concluding that a successful replication depends more on the actual effect and less on the study's precise framework.

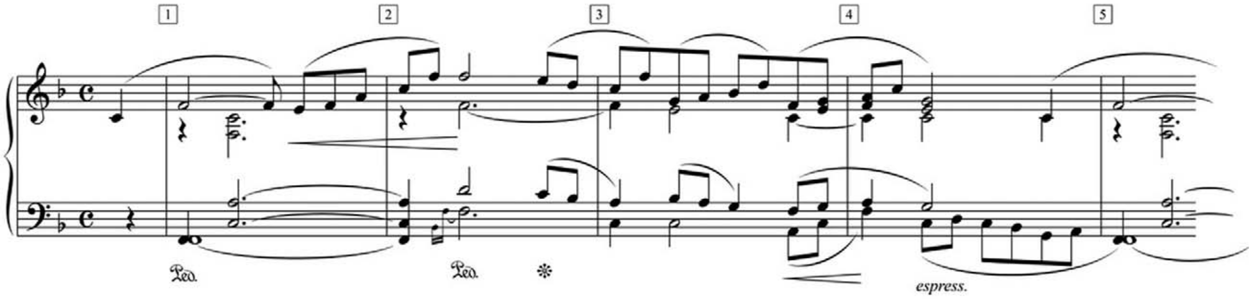

FIGURE 2. Score section of "Träumerei," Op. 15, No. 7 by Robert Schumann used in the replication study.

We would like to contribute to a tradition of replication studies in music psychology as initiated by Fischinger (2013) and theoretically outlined in the same journal issue by Frieler et al. (2013). We are convinced that in the long-term, replication studies should become a permanent element of publications in empirical music research. Third, we suggest that, as an extension of the original study, the MDH should be tested in a cross-cultural experiment to extend the initial findings and investigate their possible cross-cultural generalizability.

Our empirical hypotheses follow the results of Repp (1997) as seen in Figure 1. However, to decrease the time span of the experiment, we decided not to use the full set list of the originial study. Instead, we only tested the arithmetically "average" performance (AP), the "best" (P10) and "worst" (P8) individual performances and the median or "medium" performance (P7). We hypothesized that the best individual performance (P10) and the average performance (AP) would be rated better than a medium performance (P7), which again would be rated better than the lowest-rated performance (P8). According to Figure 1, with the exception of the "best" (P10) and "average" (AP) performances, none of the drawn standard errors of the four chosen performances overlap, resulting in three interpretations (P10, P7, P8) that represent three groups of performance evaluations plus the average performance (AP).

Method

DESIGN

Following the two hypotheses, our study used a 2 × 4 mixed methods design with the between-factor *Region* representing the two countries of data collection, Germany and Taiwan, and the within-factor *Version* representing the four interpretations of Robert Schumann's Träumerei (Op. 15, No. 7). All statistical analyses were conducted with the R Project for Statistical Computing (R Core Team, 2017); the Bayesian analyses were conducted using JASP (JASP Team, 2017).

MATERIALS

We used the same MIDI files of Schumann's "Träumerei" as Repp did (1997).[1] However, due to the unavailability of the original audio files (only a down-sampled audio file of the AP version could be obtained) we decided to reconstruct the sound files based on the Bösendorfer grand piano samples from the Vienna Symphonic Library (see the sound files Audio S1 to S4 in the Supplementary Materials that accompany the online version of this paper). Compared to the original study, all stimuli were shortened: Only the first five bars of the "Träumerei" were used (see Figure 2) and ended with a fade-out in the middle of bar 5, just at the beginning of the first reprise. The duration of the four stimuli lay between 21 and 29 seconds.

The questionnaire contained the same items for the evaluation of the interpretations as Repp (*tempo, dynamics, expression, individual*) and we added five researcher-developed items on the interpretation's conventionality: namely, whether it reminds the listener of other recordings; whether many musicians exist who interpret the piece like this or not; whether the overall quality is convincing; and whether the participant wishes to continue listening to this interpretation (Platz & Kopiez, 2013; for all items, see Table S1 in the Supplementary Materials section online). The evaluation questionnaire was composed in German and translated into Mandarin.

A PRIORI POWER ANALYSIS

We conducted an a priori power analysis (Ellis, 2010) for various reasons: First of all, we were able to directly

[1] The MIDI sound files of the original study are still available from the website http://www.haskins.yale.edu/MISC/REPP/AP1.html

derive effect sizes from the results of Repp's original study (1997), allowing a well-informed estimation of the required number of participants. As no standard errors for the "Semifinals" were reported, we used the means and standard errors for the "Finals" although these were originally calculated from the second exposure of the participants to the stimuli. By use of the software *DataThief III* (Tummers, 2006), we took estimated readings of the central tendency and dispersion of the chosen performances P10, AP, P07, and P08. The second smallest effect size we calculated from the reconstructed data was the difference between the average and the medium performance, which resulted in a medium effect of Cohen's $d = 0.5$. We did not take the smallest effect size between the average and the best individual performance as a basis because standard errors between these two version markedly overlap. This clear overlap indicates that both interpretations were rated equally well. We therefore assumed that participants in the replication study would also rate them equally or similarly and that we would not find significantly different evaluations between these performances.

Second, this information allowed us to create a statistical scenario of our experiment and calculate the minimum number of participants when taking the effect size, Type I and Type II error into consideration. A typical value for Type I error would be $\alpha = .05$, for Type II error $\beta = .20$ (Ellis, 2010).

Third, we set the upper limit of the Type II error lower than usual to aim towards a null effect, as we hypothesized that we would see no difference in ratings of German and Taiwanese participants, thus assuming no cultural difference in aesthetic evaluation. While it is impossible to directly provide evidence for a null effect, this decision would enable a meaningful effect to emerge if it existed. Also, this produces a precise estimate (e.g., a small confidence interval) of the true differences in the evaluation of the four performances with a sufficient number of participants.

Taken all together, our scenario included a medium effect size of $f = 0.25$ (which is equivalent to $d = 0.50$ or $\eta_p^2 = .06$; see Ellis, 2010, p. 41), a Type I error of $\alpha = .05$, a Type II error of $\beta = .05$ (and a test power of $1-\beta = .95$), two between-subjects groups of Taiwanese and German participants, four within-subjects measurements of the four performances, and an estimated correlation of $r = .50$ between measurements. The software G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) provided information which led to the desirable number of $N = 132$ participants.

PARTICIPANTS

The announcement of the online experiment was distributed via social media, mailing lists of university seminars (students of psychology, musicology, and education), and by word of mouth. Thus, participants in this study formed a convenience sample. However, to avoid sample bias, recruitment of participants was not limited to aficionados and specialists of classical music. The number of participants who finished the online questionnaire was $N = 260$. We discarded the responses of those participants who needed more than one hour to finish the questionnaire ($n = 17$), replied to the questions too quickly ($n = 6$), or correctly identified none or only one of the five stimuli in the reliability test ($n = 32$), in other words, at less than chance level. Of the remaining $N = 205$ participants, 59 (28.8%) detected two stimuli, 47 (22.9%) detected three stimuli, 55 (26.8%) detected four stimuli, and 44 (21.5%) detected all five stimuli correctly.

This final selection of $N = 205$ participants were 117 women and 88 men: 84 came from Germany, of whom 36 had studied or were studying music, and 121 came from Taiwan, of whom 28 had studied or were studying music. The mean age of all participants was $M = 29.4$ years ($SD = 8.9$ years); their mean score in the General Musical Sophistication Score from the Goldsmiths Musical Sophistication Index was $M = 83.2$ ($SD = 22.8$; theoretical range: 18–126; German: $M = 84.8$, $SD = 24.2$; Taiwanese: $M = 82.1$, $SD = 21.7$), which corresponds to the 75% percentile for the underlying German norm and a 53% percentile for the Taiwanese norm (see Müllensiefen, Gingras, Musil, & Stewart, 2014; for the German version, see Schaal, Bauer, & Müllensiefen, 2014; for the Taiwanese version, see Lin, Kopiez, Müllensiefen, & Wolf, 2018). Overall, $n = 130$ participants (63%; 68% of Taiwanese and 57% of Germans) had taken piano lessons at some point in their lives. The mean duration of their lessons was $M = 8.3$ years ($SD = 5.2$ years; Taiwanese: $M = 8.0$, $SD = 4.8$; German: $M = 8.8$, $SD = 5.7$). The participants' preference for classical music was $M = 80.2$ (on a scale from 0–100, $SD = 20.6$; Taiwanese: $M = 80.3$, $SD = 21.2$; German: $M = 80.0$, $SD = 19.9$); the preference for classical piano music had a slightly lower mean of $M = 76.8$ (on a scale from 0–100, $SD = 21.0$; Taiwanese: $M = 78.7$, $SD = 21.0$; German: $M = 74.0$, $SD = 20.7$).

PROCEDURE

The data collection was conducted by means of an Internet experiment (Reips, 2012), which was online from May 2 to July 15, 2016. Participants were welcomed to the study and informed about their rights as participants before giving their informed consent. We asked about their age, sex, level of education, and their musical sophistication according to the general factor of the Goldsmiths Musical Sophistication Index. They entered their preference for classical music in general and classical piano music. The next page consisted of a sound test to adjust the audio volume of the computer. Then, on the next page, the presentations of the interpretations started, and participants were informed that their subjective impression was what was relevant and that no right or wrong answers existed. Participants first familiarized themselves with the feature space by listening to the four interpretations and answering cover questions (familiarity with the piece, felt emotions, memories of and associations with the piece); interpretations and cover questions were each presented in a randomized order, respectively.

The test phase was presented as a competition of five pianists (see Supplementary Materials that accompany the online version of this paper for the instructions used in the internet survey). Participants evaluated the five performances, presented in a randomized order, using the nine items from Table S1 in the Supplemental Materials section online. The five performances consisted of the three individual performances P10, P07, P08 and the average performance AP, which was retested for reliability purposes. Each performance was played only once. After each evaluation, participants listened to short snippets (8–12 s long, from the beginning to the second beat in the second measure; see Figure 2 and sound files Audio S5 to S8 in the Supplementary Materials) of the four versions at the bottom of the website and decided which interpretation they had just listened to and evaluated. This detection task was the basis for the correct identification of the stimuli, which was used to select only the attentive participants. On the last page of the questionnaire, participants were thanked and given the possibility to enter their own feedback. The mean overall duration of this online study was $M = 15.7$ min ($SD = 3.92$).

Results

EXPLORATORY FACTOR ANALYSIS

First, we performed a maximum-likelihood exploratory factor analysis (R function *fa* from the *psych* package, Revelle, 2017) with oblimin rotation on the nine items (see Table 1). We chose an oblique method of rotation because we assumed that even if more than one factor appeared, the different constructs represented by the factors and items would still be correlated to some degree (Abell, Springer, & Kamata, 2009; Field, Miles, & Field, 2012). The Kaiser-Meyer-Olkin (KMO) factor adequacy was larger than .50 for each item, and Bartlett's test for sphericity was significant, $\chi^2(36) = 3486.43$, $p < .001$. Following the Kaiser criterion and the scree, one factor was sufficient to describe the latent variable structure in the data (eigenvalue of a second factor = .92). As only one factor was extracted, there was also no risk of over-extraction of factors (Costello & Osbourne, 2005). The Root Mean Square Error of Approximation index was $RMSEA = 0.03$ and therefore better than a typical cutoff of 0.06. The Bayes Information Criterion was $BIC = 621.15$. The Root Mean Square Residual was $RMR = 0.134$. Finally, the Comparative Fit Index was $CFI = .77$. Theses fit indices show a moderate to good fit of the model for the data and allow further statistical investigation.

TABLE 1. *Results of the Exploratory Factor Analysis*

Item	Factor 1
Overall quality	.86
Expression	.82
Dynamics	.73
Tempo	.66
Continuing to listen	.63
Explained variance	36.2%
Eigenvalue	3.42

After removing the item *other recordings*, which reduced Cronbachs α to .83, Factor 1 consisted of the items *tempo, dynamics, expression, overall quality*, and *continuing to listen* (Cronbachs $\alpha = .86$). The variance accounted for by this model was 36.2%. Factor scores were calculated using a regression method and used as dependent variables in the comparison of the musical evaluations and in the analysis of variance. These factor scores followed a distribution with $M = 0.00$ and $SD = 0.94$. Due to the different scale of items (mix of 4-point-rating scale and dichotomous items), factor scores were used instead of a mere sum score of the items. Factor 1 can be interpreted as Musical Quality, which encompassed items on the appropriateness of tempo, expression, and dynamics as well as the overall quality and the participant's wish to continue to listen to this interpretation.

Probabilistic Analysis of the Unidimensional Structure of the Factor Musical Quality. In addition to the exploratory factor analysis, we wanted to confirm the unidimensional structure of the factor that was found. To the best of our knowledge, the best suitable analysis for this purpose is item response theory (IRT; see Bond & Fox, 2007; Wilson, 2005), which uses a probabilistic approach with detailed information on item characteristics and which exceeds the focus on error-free measurement of classical test theory. We used quasi-exact statistical tests that evaluate: 1) the local independence within the model, 2) its homogeneity (or unidimensionality), 3) the quality of the item characteristic curves, and 4) the specific objectivity of the item set. The general idea behind these properties are as follows: 1) local independence is achieved if the answer to one item is not dependent on the answer to another item; 2) Items are homogenous if they do not belong to other latent dimensions than the majority of items; 3) item characteristic curves should follow the Logit function, which is parallel and strictly increasing and shows a continuously rising answer probability without any sudden leaps; (4) lastly, it does not matter which exact set of items is used to measure a participant's score on the latent variable—unlike in classical test theory, where only the whole scale can be used for measuring a construct. After these considerations, we decided to recode the items to a dichotomous scale (two lower points 1 & 2 to 0; two higher scale points 3 & 4 to 1) and benefit from the higher power of this quasi-exact test family (Koller, Maier, & Hatzinger, 2015), compared to the standard parametric tests within IRT, rather than use the theoretically more constrained rating scale model. Ultimately, our precise item selection after the EFA did not fulfil the condition of the rating scale model either way, as scale types were different between items (mix of 4-point rating scale and dichotomous items). Therefore, the items had to be recoded.

The four qualities of the test model mentioned above were necessary to find out whether participants' replies depended solely on their evaluation of the musical quality of the piece. If the evaluation was also influenced by other matters, the data would fall short on one or several of these qualities. (For a more in-depth discussion on the theoretical and statistical features of these tests, see Koller & Hatzinger, 2013; Koller et al., 2015).

Local independence and homogeneity were assessed together using three different tests: The global test T11 showed a significant difference between observed and expected inter-item correlations ($p < .001$) and required more detailed testing using the test T1. Here, some item correlations were significantly too high ($p < .01$)—and therefore a threat to the items' local independence—between the item *quality* and the items *dynamics*, *expression,* and *continue to listen*, respectively. These results allowed the conclusion that the item *quality* was dependent on other items, which does not come as a surprise: In terms of its content, this item is most similar to the general construct and already shows in itself a consolidation of the other items. Next, reduced inter-item correlations in the test T1m—and therefore a threat to the items' homogeneity—were found between the item *tempo* and the items *dynamics*, *expression,* and *continue to listen*, as well as between *dynamics* and *continue to listen*. However, using the median as a split-criterion in test Tmd concerning violations of homogeneity between the easier-to-agree and harder-to-agree items, we found no significant differences ($p > .19$). The empirical item characteristics curves were inspected and showed small deviations for the item *tempo*.

Finally, the specific objectivity of the items was tested using the tests T10 and T4. The global test T10 split the participants according to their median into a more "agreeing" and a more "disagreeing" group. No difference in response behavior was found. The test T4 focused on each item and on whether it was too easy to agree to for the more "disagreeing" group. As can be suspected from the previous item characteristic curve results, the item *tempo* showed significantly different results in test T4 as well.

If this scale development and item selection procedure had had the aim of building a stable set of items for repeated use with various participant samples, we would have started with a larger item set to begin with and would now argue for dismissing the item *tempo*. However, as our aim was to replicate a well-known and highly relevant effect of music perception and evaluation, we decided to keep the critical item for the calculation of the score of the latent variable. This entire scale development procedure served the primary goal of selecting the best items, retaining a sensible and large enough set of items, and revealing the resilience of the developed scale. Pragmatically, one could argue that this scale was suitable for our analysis with and without the item *tempo* as the following ANOVA was repeated without *tempo,* and the results did not change.

COMPARISON OF THE MUSICAL EVALUATIONS

A mixed ANOVA with one between factor (Region) and one within factor (Version) and their interaction factor was calculated to test for group differences in the evaluation of the dependent variable Musical Quality of the four interpretations (for descriptive statistics see Table 2 and Figure 3).

There was a significant main effect for Version, $F(3, 609) = 29.49$, $p < .001$, $\eta_p^2 = .13$, and no significant effects for either Region, $F(1, 203) = 0.05$, $p = .82$ (see also Figure 4) or the interaction between Region and Version, $F(3, 609) = 0.27$, $p = .84$. A post hoc Tukey test showed that all interpretations were rated differently

TABLE 2. ***Descriptive Statistics of the Latent Factor Musical Quality***

	Musical Quality Rating		
	M	*SD*	*95% CI*
AP	0.26	0.80	[0.15, 0.37]
AP Retest	0.32	0.79	[0.21, 0.43]
P10	−0.03	0.89	[−0.15, 0.09]
P7	−0.01	0.91	[−0.14, 0.12]
P8	−0.54	1.05	[−0.68, 0.40]

Note. Musical Quality Rating is a factor score (calculated by a regression method).

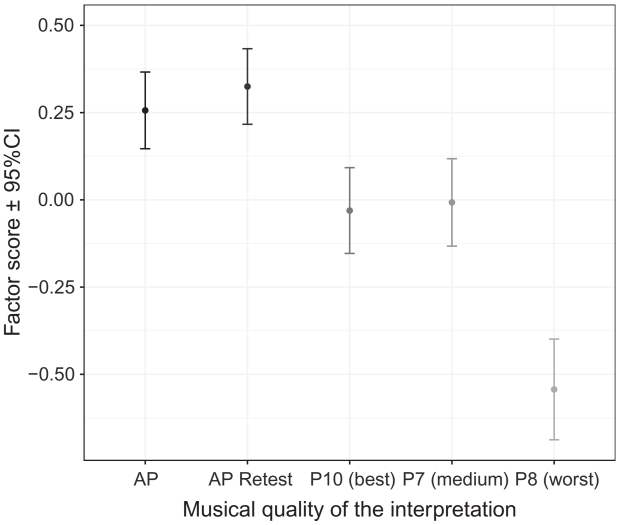

FIGURE 3. Overall ratings of Musical Quality (factor scores) of the four versions. Positive values indicate better evaluation.

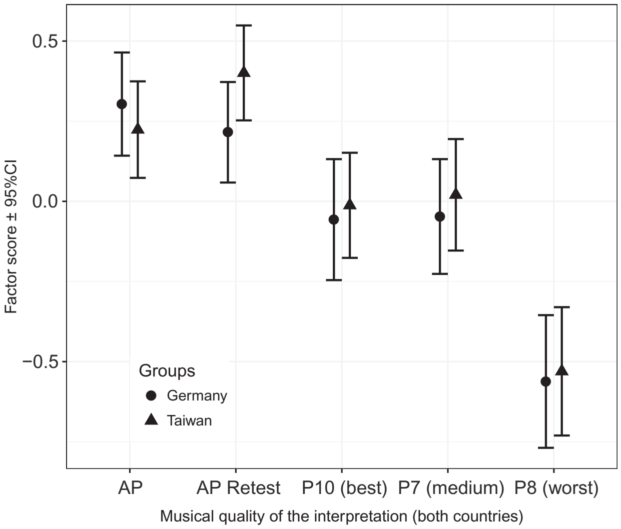

FIGURE 4. Overall ratings of Musical Quality (factor scores) of the four versions for the respective German and Taiwanese subsample. Positive values indicate better evaluation.

from each other pairwise with regards to their Musical Quality ($p < .01$)—except for interpretations P10 and P7, the interpretations chosen as "best" and "medium" from Repp's original study.

RESULTS FROM BAYESIAN ANALYSIS

For the purpose of model comparison, we calculated the Bayes Factor (Kruschke, 2015) between the null model and all other models (one for the within-subject factor Version, one for the influence of the between-subjects effect Region, one for both factors and, finally, one for the influence of both factors and the interaction between Version and Region on the performance evaluation). Thus, we conducted a Bayesian ANOVA (Rouder, Morey, Speckman, & Province, 2012) with equal prior model probabilities as default prior scales (see Table 3, column P(M)) with the software JASP (JASP Team, 2017). The analysis revealed "decisive" evidence for the preference of the first main effect model (Version) against the null model by a Bayes factor of 3.33e+15 (see column "BF_{10}" in Table 3; for a classification of Bayesian Factors (BF), see Jeffreys, 1969). This means that the main model was 3.33e+15 times more likely than the null model. However, neither Region nor the inclusion of the interaction of Version and Region (last model) showed a substantial model improvement compared with the null model ($BF_{10} = 0.10$, i.e., 1/0.10 = 10.53 in favor of the null model). Additionally, the first model (Version) was also strongly preferred over the two-main effects model ($BF = 10.12$). Conclusively, the first model based on Version as the only main effect showed the highest preference compared with all other models, as well as the highest posterior model probability, which could be the prior probability for future replications (for the interpretation of the *BF* outcome parameters and benchmarks, see Wagenmakers, Love, et al., 2017).

Discussion

This study successfully replicated Repp's (1997) results on the minimal-distance hypothesis that people prefer an averaged musical interpretation to individual interpretations. In contrast to Repp's findings, the main result of the current study was the comparatively lower rating of the "best" version (P10), which was rated equally as the "medium" interpretation (P7). Overall, this main result provides evidence for a stability in human perception and a general preference for averaged versions—a phenomenon that was clearly supported by the culturally universal evaluations among the Taiwanese and German participants. This last

TABLE 3. *Comparison of Bayes Models*

Model Comparison					
Models	*P(M)*	*P(M|data)*	BF_M	BF_{10}	error %
Null model (incl. subject)	0.200	2.729e-16	1.092e-15	1.000	
Version	0.200	0.909	39.874	3.330e+15	0.577
Region	0.200	2.589e-17	1.036e-16	0.095	1.511
Version + Region	0.200	0.090	0.395	3.290e+14	1.197
Version + Region + Version * Region	0.200	0.001	0.005	5.029e+12	2.009

Note. All models include subject; *P(M)* = prior model probability; *P(M|data)* = posterior model probability; BF_M = change from prior to posterior model odds; BF_{10} = Bayes Factor in favor of each model compared with the null model (see JASP Team, 2017: for details see Wagenmakers, Love, et al., 2017; Wagenmakers, Marsman, et al., 2017).

finding, which supports the null hypothesis, was probably not a result of a low test power: With the current sample size, the probability of detecting a medium effect size was 95% in our implementing the study. Rather, it was either a result of the cultural universality of the MDH or evidence of a high cultural similarity between Taiwanese and German people in their relationship to classical music. However, considering the study of Kopiez et al. (1999), in which Ghanaian and German participants rated the averaged version best, we currently assume cultural universality of the MDH in music. Transferring the minimal-distance hypothesis to the visual modality and, more specifically, the attractiveness of faces, the meta-analysis by Langlois et al. (2000) combined 82 studies on adults and provided evidence that people within and across cultures rate the attractiveness of faces highly similarly. However, to extend the empirical basis for the MDH, future studies should consider classical pieces that are less popular and of different character than the "Träumerei" as well as music from other genres.

In what way is this study an extension of Repp's original work (1997)? First, our sample was not drawn from a population of experts in classical music but was a convenience sample of the general population. This is confirmed by the musical sophistication of the participants, which lay slightly above the average range of sophistication scores for the German sample (75% percentile) and close to the mean of the norm distribution for the Taiwanese sample (53% percentile, see Lin et al., 2018). Second, due to the general advancement of the experimental method over the last two decades, we have been able to report results in more detail and allow statistically more precise conclusions from confidence intervals in addition to the less precise readings from error bar diagrams. Lastly, the original results only relied on the answers to one single item, the overall *musical quality* (measured on a scale from 0 to 10). We employed a statistically mixed approach and validated our evaluation scale by classical means using an exploratory factor analysis and a more advanced method using item response theory. Therefore, we have obtained a more stable construct, which is not without its flaws–particularly regarding the possibly non-fitting *tempo* item–but provides more valid and detailed results.

In light of the current replication crisis, these highly similar results between the original and the replication study seem to be more of an exception than the rule (Open Science Collaboration, 2015). Repp's original results seem to hold steady and stand in line with the successful replications of music psychological results, such as Deutsch's octave illusion (1974), replicated by Oehler and Reuter (2013), and the effect of playing from memory on performance evaluation (original study by Williamon, 1999; replicated by Kopiez, Wolf, & Platz, 2017). In our opinion, the most probable reason for the replication of the effect at hand lies in its multisensory feature. This is not merely an effect in aesthetical perception in music but was also investigated for human faces (Langlois & Roggman, 1990) as well as birds, fish, and automobiles (Halberstadt & Rhodes, 2003): For the last three specimen, the authors again found stable effects of averaging, but also investigated the role of familiarity and found that "both [averageness and familiarity] are linked to the distribution of traits that is perceived as attractive for whatever evolutionary, sociological, or cultural reasons" (p. 155). This conclusion confirms the relevance of more research on this interaction, which should lead to interesting questions about music evaluations. As the degree of familiarity with a specific style certainly influences evaluations, an experiment with an unfamiliar musical style could shed some light on the interactions between averageness and familiarity for musical stimuli. Outside the realm of classical music, these results could also be relevant for (digital) editing processes in general. When a higher attractiveness of a product is targeted, less extreme and more average versions might lead to more success. On the other hand, such processes are repeatedly counteracted when the outcomes become too well-known, unimaginative, or predictable.

This study has shown that the preference for an averaged musical interpretation still exists after decades of quantized and computer-controlled music. Future investigators on the MDH in music are faced with the task of producing several interpretations of a musical piece and developing a new procedure for averaging them. With digital audio recording techniques and MIDI interfaces, this task is certainly easier than in the past, but still requires a sophisticated routine.

Of course, our study had to face some limitations, starting with the selection of musical pieces. However, we did not extend this for several reasons: First, we wanted to find out whether the preference for an averaged version still existed after all the years. Second, if we had investigated another piece of music (using a within-subjects design), the duration of the study could have resulted in a raised risk of an increased drop-out rate. As reported by Vicente and Reis (2010), the length of questionnaires in online studies has a significant influence on nonresponse rates: After about 10 minutes of survey duration, the drop-out likelihood increases significantly (Crawford, Couper, & Lamias, 2001), starting from an automatic drop-out rate of 10%, which seems to be characteristic of internet experiments (Hoerger, 2010). A maximum duration of 15 to 30 minutes seems to be the guideline (Vicente & Reis, 2010). Third, the production of the average version was not fully documented in Repp (1997) and could therefore not be accurately repeated – we would have compared two versions of averaging and would not have been able to unambiguously ascribe potential differences in evaluation to the original and another piece. Fourth, we followed the frequent recommendation to first replicate studies in a similar manner to the original study (see, e.g., Frank & Saxe, 2012; Frieler et al., 2013, p. 268; Ioannidis, 2014, p. 4). Further research using different pieces of music from different genres and with different instruments would shed new light on the MDH and its generalizability. For now, we can assume that this effect will be repeated with music of different characteristics because it has been proven as a reliable phenomenon in different modalities.

A final limitation concerns the role of extremes in aesthetic judgements, and we cannot exclude that aesthetic appreciation is not only caused by mainstream interpretations. For example, research from evolutionary biology showed that Symons' (1979, p. 197) "averageness hypothesis" might be an oversimplified explanation: Perrett, May, and Yoshikawa (1994) observed that the 15 top-rated female faces outperformed the total average of 60 portraits. Furthermore, a caricature version based on an increased contrast of features (e.g., lip-nose distance, eye distance) from the grand average and from the top-15 average versions again outperformed all other average versions. In other words, at least for facial attractiveness, the perception of beauty might rely on the averageness effect but also on more sophisticated mechanisms, such as the increased contrasts of features. In the domain of music, there is at least one famous episode when an interpretation that was far beyond the mainstream performance caused significant aesthetic interest: In 1962 the pianist Glenn Gould gave a performance of Brahms' Piano Concert in d Minor (Op. 15) under the direction of Leonard Bernstein, that was so unconventionally slow that Bernstein distanced himself from it in a short introduction (Page, 1984, pp. 70-71). However, Gould's intention was not to create a caricature of the piece but to bring in some fresh ideas on the interwoven motivic relationship between the solo piano and orchestra accompaniment. From Gould's point of view, this new approach to the Brahms concerto required a significant slowing down of the tempo to reduce the contrasts between the two musical actors. Although this might appear eccentric or mannered, critics were predominantly enthusiastic about Gould's radically new beginning: "Elements nobody previously had paid much attention to suddenly sprang in high relief. (. . .) The music was passing through a mind that took nothing for granted." (Schonberg, 1987, p. 481)

The present study has successfully replicated the results of the minimal-distance hypothesis investigated by Repp (1997), although we have significantly extended the evaluation questionnaire, updated the stimuli material, and modernized the statistical procedures. The evaluations of participants from Germany and Taiwan were compared, and they support the conclusion that the preference for average musical stimuli compared to individual interpretations—in other words, the MDH—seems to be culturally invariant. However, the participants' experience with, and preference for, classical (piano) music was comparable and presumably above average. Stronger evidence for cultural universality would be provided by more controlled and representative samples from a wider range of musical characteristics such as tempo variance or dynamics.

On a meta level, and in light of the current replication crisis in psychology, this positive result is encouraging for further replication studies in music psychology. However, there remains still one question: What should the performer do? Should he or she play a highly individualistic interpretation or, instead, not deviate too far from the norm? As stated by Repp in the original study, the art of interpretation is rather a dynamic process: "A pleasing, 'prototypical' performance is one thing; an interesting, individual performance is another. Competing centripetal and centrifugal forces keep classical music performance alive!" (B. Repp, personal communication, July 26, 2017).

Author Note

Correspondence concerning this article should be addressed to Anna Wolf, Institute for Systematic Musicology, University of Hamburg, Neue Rabenstraße 13, 20354 Hamburg Germany. E-mail: anna.wolf@uni-hamburg.de

References

Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments.* New York: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Costello, A. B., & Osbourne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research Evaluation*, *10*(7), 1–9. Available online: http://pareonline.net/getvn.asp?v=10&n=7

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys. *Social Science Computer Review, 19*, 146–162. DOI: 10.1177/089443930101900202

Cumming, G. (2014). The new statistics. *Psychological Science*, *25*, 7–29. DOI: 10.1177/0956797613504966

Deutsch, D. (1974). An auditory illusion. *Nature*, *251*(5473), 307–309. DOI: 10.1038/251307a0

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge, UK: Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Los Angeles, CA: Sage.

Fischinger, T. (Ed.). (2013). Replication in music psychology [Special issue]. *Musicae Scientiae*, *17*, 263-264. DOI: 10.1177/1029864913502763

Frank, M. C. & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*, 600–604. DOI: 10.1177/1745691612460686

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K. B., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae*, *17*, 265–276. DOI: 10.1177/1029864913495404

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single resultant figure. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *8*, 132–144. DOI: 10.2307/2841021

Halberstadt, J., & Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin and Review*, *10*, 149–156. DOI: 10.3758/BF03196479

Hoerger, M. (2010). Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking, 13*, 697–700. DOI: 10.1089/cyber.2009.0445

Ioannidis, J. P. A. (2014). How to make more published research true, *PLOS Medicine, 11*, e1001747. DOI: 10.1371/journal.pmed.1001747

JASP Team (2017). *JASP* (Version 0.8.2) [Computer software]. Amsterdam: Department of Psychological Methods.

Jeffreys, H. (1961). *The theory of probability.* Oxford, UK: Clarendon Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. DOI: 10.1027/1864-9335/a000178

Koller, I., & Hatzinger, R. (2013). Nonparametric tests for the Rasch model: Explanation, development, and application of quasi-exact tests for small samples. *InterStat*, *2.*

Koller, I., Maier, M. J., & Hatzinger, R. (2015). An empirical power analysis of quasi-exact tests for the Rasch model: Measurement invariance in small samples. *Methodology*, *11*, 45–54. DOI: 10.1027/1614-2241/a000090

Kopiez, R., Langner, J., & Steinhagen, P. (1999). Afrikanische Trommler (Ghana) bewerten und spielen europäische Rhythmen [African drummers (Ghana) evaluate and perform European rhythms]. *Musicae Scientiae*, *3*, 139–160.

Kopiez, R., Wolf, A., & Platz, F. (2017). Small influence of performing from memory on audience evaluation. *Empirical Musicology Review, 12.* DOI: 10.18061/emr.v12i1-2.5553

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* New York: Elsevier.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*, 390–423. DOI: 10.1037//0033-2909.126.3.390

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*, 115–121. DOI: 10.1111/j.1467-9280.1990.tb00079.x

Lin, H.-R., Kopiez, R., Müllensiefen, D., & Wolf, A. (2018). *The Chinese version of the Gold-MSI: Adaption and validation of an inventory for the measurement of musicality in a Taiwanese sample*. Manuscript in preparation.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing Musical sophistication in the general population. *PLOS ONE*, e89642. DOI: 10.1371/journal.pone.0089642

Oehler, M., & Reuter, C. (2013). The octave illusion and handedness: A replication of Deutsch's 1974 study. *Musicae Scientiae*, *17*, 277–289. DOI: 10.1177/1029864913493801

Open Science Collaboration. (2015). *Estimating the reproducibility of psychological science*. Science, *349*(6251), aac4716–aac4716. DOI: 10.1126/science.aac4716

Page, T. (Ed.) (1984). *The Glenn Gould reader.* New York: Vintage Books.

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature*, *368*(6468), 239–242. DOI: 10.1038/368239a0

Platz, F., & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musicae Scientiae*, *17*(2), 167–197. DOI: 10.1177/1029864913486369

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

Reips, U.-D. (2012). Using the Internet to collect data. In *APA handbook of research methods in psychology.* (Vol. 2, pp. 291–310). Washington, DC: American Psychological Association. DOI: 10.1037/13620-017

Repp, B. H. (1997). The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception*, *14*, 419–444. DOI: 10.2307/40285732

Revelle, W. (2017). *Package "psych"* [Electronic reference]. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rouder, J. N., Morey, R. D., Speckman, P. L., Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356–374. DOI: 10.1016/j.jmp.2012.08.001

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe [The Gold-MSI: Replication and validation of a questionnaire instrument for measuring musical sophistication, based on a German sample]. *Musicae Scientiae*, *18*(4), 423–447. DOI: 10.1177/1029864914541851

Schonberg, H. (1987). *The great pianists*. New York: Simon and Schuster.

Symons, D. (1979). *The evolution of human sexuality.* Oxford, UK: Oxford University Press.

Tummers, B. (2006). *DataThief III* [Computer software]. Retrieved from http://datathief.org

Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review, 28*(2), 251–267. DOI: 10.1177/0894439309340751

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, *25*, 58–76. DOI: 10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*, 35–57. DOI: 10.3758/s13423-017-1343-3

Williamon, A. (1999). The value of performing from memory. *Psychology of Music*, *27*(1), 84–95.

Williams, J. (1966). *Leonardo da Vinci: Abenteuer des Geistes.* Reutlingen, Germany: Ensslin & Laiblin.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New York: Psychology Press.