When the Eye Listens: A Meta-analysis of How Audio-visual Presentation Enhances the Appreciation of Music Performance

Friedrich Platz & Reinhard Kopiez
Hanover University of Music, Drama and Media, Germany

The visual component of music performance as experienced in a live concert is of central importance for the appreciation of music performance. However, up until now the influence of the visual component on the evaluation of music performance has remained unquantified in terms of effect size estimations. Based on a meta-analysis of 15 aggregated studies on audio-visual music perception (total $N = 1,298$), we calculated the average effect size of the visual component in music performance appreciation by subtracting ratings for the audio-only condition from those for the audio-visual condition. The outcome focus was on evaluation ratings such as liking, expressiveness, or overall quality of musical performances. For the first time, this study reveals an average medium effect size of 0.51 standard deviations — Cohen's *d*; 95% CI (0.42, 0.59) — for the visual component. Consequences for models of intermodal music perception and experimental planning are addressed.

Received May 17, 2011, accepted November 19, 2011.

Key words: audio-visual perception, performance evaluation, intermodal perception, music performance evaluation, meta-analysis

Why do people often prefer the experience of a live concert when they could enjoy errorless high quality recordings of the same performances in a pleasant listening environment? One explanation is that the visual component of the live music performance contributes significantly to the appreciation of music performance (Bergeron & Lopes, 2009; Cook, 2008). Audio-visual music performance experience is an important factor in many musical traditions (Frith, 1996; Vines, Krumhansl, Wanderley, Dalca, & Levitin, 2006) and cannot be discounted in our mediatized society (Auslander, 2008). This view is also supported by historical reports on famous artists, such as Franz Liszt (Burger, 1986; Schumann, 1854/1985). In summary, the attractiveness of live performances, concert recordings on DVDs, and music television give support for the continued attractiveness of audio-visual music mediation. In contrast, from the perspective of musical communication approaches, successful transmission of musical meaning lies in the acoustical communication brought about by performers and composers (Juslin, 2005).

Until recently, potential influences of visual components on the appreciation of music have been widely neglected in music perception research (Gabrielsson, 2003; for an exception see Davidson, 1993). The following list contains some exceptions that should be mentioned: Finnäs (2001) observed an increasing interest in the visual component's influence on music performance perception, and Schutz (2008, p. 90) concludes that a "variety of musical properties and types of evaluations can be affected by the visual information." However, from the perspective of models of musical communication through acoustical realization of structural features, the visual component only plays a role as a mere additive or supporting component. Against the background of current models of multisensory perception, we argue that there is no justification for the separation of modalities or the focus on the aural component in music perception only (Thompson, Graham, & Russo, 2005).

Although numerous studies show a general positive effect of the visual component on the experience and evaluation of music (McPherson & Thompson, 1998; Thompson et al., 2005), it is unclear how much visual aspects of music influence the evaluation of music performance (Bermingham, 2000). Therefore, our intention is to quantify the influence of the visual component on performance evaluation. To the best of our knowledge, this is the first attempt to summarize existing data by estimating the effect size (Humphrey, 2011).

Up until now, two methods have been used for an objective comparison of research results: the systematic

Music Perception volume 30, issue 1, pp. 71–83. issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2012.30.1.71

review and the meta-analysis. Definitions of systematic reviews (Cooper, 2010; Cooper, Hedges, & Valentine, 2009) can be understood as a selection of literature based on "pre-specified eligibility criteria in order to answer a specific research question" (Higgins & Green, 2009, p. 6). The review approach uses a reproducible methodology that aims to reveal those studies that would meet all the criteria. The second approach, meta-analysis, is the use of statistical methods to summarize the results of independent studies (Glass, 1976). However, up until today most of the study results offer only significant reports by presenting a p value ($p < .05$) corresponding to a test of significance (Borenstein, 2009; Sedlmeier, 2009). Indeed, the p value could be part of the effect size estimation as a representation of relevance, but it is also partly a function of sample size (Borenstein, 2009). According to Cohen (1988), effect size is the difference of means (M_1–M_2) divided by the pooled standard deviation (SD_{pooled}) of both random sample groups (see Equation 1).

$$d = \frac{M_1 - M_2}{SD_{pooled}} \tag{1}$$

SAMPLE SIZE, SIGNIFICANCE, EFFECT SIZE, AND TEST POWER

Experimental psychologists consider the effect size in the context of three additional parameters: sample size, significance criterion, and test power. The four parameters are interconnected. If three of these four parameters are known, the remaining parameter is fully determined. The term test power describes the probability that a statistical test will result in the conclusion that an a priori defined effect size exists, if it really exists (Cohen, 1988 p. 4). In general, researchers should run a power analysis before they conduct an experiment (Faul, Erdfelder, Buchner, & Lang, 2009). Unfortunately, there is a dilemma in knowing at least three of the four parameters. They have to be estimated on the basis of a posteriori power analyses of previous similar studies (Ellis, 2010). Often the main problem of reconstructing those parameters lies in the lack of necessary information on descriptive statistics, especially for effect size calculations (Cortina & Nouri, 2000). In case of insufficient information, effect sizes can be estimated (Seifert, 1991).

For the purpose of a meta-analytic approach, studies' effect sizes have to be weighted before they are aggregated. The weight of every study's effect size reflects its degree of precision depending on its variation as a function of sample size (Ellis, 2010). Consequently, studies with smaller sample sizes, combined with a greater variation, will result in a smaller weight compared with studies containing larger sample sizes and smaller variations. These individual studies' weights were used as precision estimators and can clearly differ from each other. Differences in weights can lead to statistical heterogeneity. The problem of weighting is encountered in the aggregation of studies with two different models: on the one hand, the so-called fixed-effect model and on the other hand the so-called random-effects model (Cooper et al., 2009). Finally, the result of a meta-analysis is the weighted mean effect size of all included studies. Compared with an individual study's effect size, it reflects more precisely a data point as well as an interval estimation of the effect size in the population (Ellis, 2010, p. 95). Moreover, a meta-analysis generally increases statistical power, whereas the standard error of the weighted average effect size will be reduced at the same time (Cohn & Becker, 2003).

All these meta-analyses share two common goals: First, they can postulate an interval of effect size estimation in a population based on aggregated effect sizes of individual studies. Second, based on the results, meta-analytic studies can give an evidence-based answer to those questions that reviews or replication studies cannot give; for example, due to their collection of significant and insignificant results. In previous studies, meta-analysis has been successfully applied to various topics related to the field of music cognition (Chabris, 1999; Hetland, 2000; Kämpfe, Sedlmeier, & Renkewitz, 2010; Pietschnig, Voracek, & Formann, 2010), and has been shown to be an important procedure for the production of verified knowledge.

RATIONALE OF THE STUDY

The aim of our study was two-fold: First, we identified all relevant publications by means of a systematic literature review to answer the question of how strongly the visual component influences the evaluation of music performances. Second, we wanted to quantify the effect of the presentation mode of a music performance on the audience's evaluation. For the first time, this meta-analysis was supposed to reveal the visual component's effect size in audio-visual music performance evaluation.

Method

SAMPLE OF STUDIES

Our systematic review started with a query for literature in the most comprehensive electronic bibliographic databases PsycINFO, ProQuest, PubMed, RILM, ISI-Web of Knowledge, and DOAJ. The query was conducted from October to December 2010 and considered publications from 1940 to the beginning of 2011. According to Ellis (2010), meta-analytic approaches must be aware of any over- or under-estimation of total effect size. A first step to address

such estimation errors is the inclusion of publication types such as dissertations or conference proceedings (Rothstein, Sutton, & Borenstein, 2005). The reason for the inclusion of "grey" literature is that nonsignificant studies often remain unpublished, also known as the "file drawer problem" (Rosenthal, 1979). The danger of overestimating the final effect size increases if only peer reviewed journal articles are taken into account. To avoid language and publication bias (Rothstein et al., 2005), publications in English and German have been considered as well as dissertations and conference proceedings. Due to the non-availability of these two sources in the main databases, both sources are often assigned to the "grey literature problem" (Cooper et al., 2009).

First, we wanted to identify all types of publications that shared at least the combination of the two keywords "music" and "audio-visual." Consequently, we used the combination of the keywords "music AND (audio-visual OR audiovisual)" as search criteria. As a first partial result, Table 1 shows the number of suggested publications in relation to the combination of search terms for every database. In the next step, all studies that did not meet our criteria of music performance evaluation were removed from this suggested publication sample. Studies that were included had to consider ratings of overall impression, liking, expressiveness, or overall quality as at least one dependent variable.

Finally, we excluded all studies in which artificially rendered combinations of audio and visual stimuli were used, such as point-light technique, pictures, trick animations, or movie sequences. These stimuli may be less biologically plausible, leading to extraneous perceptual effects. As known from studies in neurocognition, incongruent audio-visual stimuli evoke a late positive ERP (P2), which is an indicator of a higher cognitive effort in perceptual processing (Spreckelmeyer, Kutas, Urbach, Altenmüller, & Münte, 2006; Stekelenburg & Vroomen, 2007). As Beauchamp, Lee, Argall, and Martin (2004) could show, the level of dynamic synchrony between audio and visual processing channels will be higher if both audio-only and audio-visual stimuli come from the same source. In other words, biologically plausible and congruent audio-visual stimuli may lead to low-level automatic integration. With regard to our selection criteria, our query revealed a corpus of 33 studies (see Appendixes B and C).

SELECTION CRITERIA FOR META-ANALYSIS

Studies were included in the meta-analysis as long as they complied with the following criteria: (a) hypothesis-testing design; (b) no continuous (physiological) data; (c) presentation mode realized with at least two conditions (audio vs. audio-visual), with the audio presentation as the control condition; (d) usage of items such as "liking," "expressiveness," "overall quality," or "overall impression" for performance evaluation; (e) report of minimal statistical information (e.g., mean, standard deviation, sample size); (f) report of study design in case of ANOVA methods; (g) indication of correlation for the within subjects factors in case of a repeated measures design; (h) report of p, t, or F values for the estimation of effect sizes when no descriptive statistics were given; (i) only one study in case of publication of multiple studies based on the same dataset. Additionally, we tried to retrieve missing statistical information by establishing personal communication with the author(s). Finally, we included 15 studies (Appendix B) and excluded 18 studies (Appendix C). Moreover, to control for the methodological features of the 15 studies, a coding sheet was used (Valentine & Cooper, 2008, see Appendixes D and E).

ESTIMATION/APPROXIMATION OF EFFECT SIZES

If no descriptive statistical values were reported, but F, t, and p values and sample size were given, Cohen's d was estimated (Borenstein, 2009; Cortina & Nouri, 2000). Additionally, in case of incomplete statistical information, the effect direction was extracted from written reports in the studies' results sections. If a nonsignificant result was reported, but no t or F value given, we used $p = .06$ (one-tailed tests) in combination with group sample size and an estimation of the respective cumulative critical statistical value for the approximation of d. Statistical calculations including effect size estimation and aggregation were conducted by the software Comprehensive Meta-Analysis (Borenstein, 2010).

DATA AGGREGATION

The study-related effect sizes were aggregated using methods of random-effects model calculations (Cooper et al., 2009; see also Appendixes F and G). Although the test for heterogeneity of effect size parameters (Q) failed the critical value of $\alpha = .10$, we rejected the assumption of homogeneity of effect sizes (H_0) as a result of an a priori power analysis. In the case of a nonsignificant Q value based on $k = 15$ studies (Appendix G), the power of homogeneity tests would not be sufficient ($1-\beta < .90$) to reject the assumption of heterogeneity (H_1) (Hedges & Pigott, 2001).

TEST FOR PUBLICATION BIAS

The documentation of research projects in reviewed journals reflects only a part of research activities due to a selective decision process. Therefore, it may be suggested that unpublished literature differs systematically from

TABLE 1. **Overview of Search Strategies Used to Reveal All Listed Literature Regarding Evaluation Differences of Music Performance Depending on Presentation Mode.**

Literature search			Systematic review	
Database	Search strategy	Number of studies suggested	Number and proportion of studies included	Authors of studies included
PsycInfo	((audiovisual OR audio-visual) AND music*).mp.	132	7 (21.21%)	Ryan & Costa-Giomi (2004), Ryan et al. (2006), Wapnick et al. (2009), Wapnick et al. (1997), Wapnick et al. (1998), Wapnick et al. (2000), Wapnick et al. (2004)
ProQuest	cabs(music* AND (audiovisual OR audio-visual)	109	4 (12.12%)	Howard (2009), Min (2001), Siddel-Strebel (2007), Zumpella (1993)
PubMed	music* AND (audiovisual OR audio-visual)	82	0	
RILM	music AND (audiovisual OR audio-visual)	729	13 (39.39%)	Bullerjahn & Lehmann (1989), Howard (2009), Min (2001), Peddell (2004), Ryan & Costa-Giomi (2004), Ryan et al. (2006), Schmidt (1976), Siddel-Strebel (2007), Wapnick (2009), Wapnick et al. (1998), Wapnick et al. (2000), Wapnick et al. (2004), Zumpella (1993)
ISI Web of Knowledge	TI = (audio* AND (audiovisual OR audio-visual)) Refined by: Web of Science Categories = (EDUCATION EDUCATIONAL RESEARCH OR PSYCHOLOGY EXPERIMENTAL OR PSYCHOLOGY DEVELOPMENTAL OR PSYCHOLOGY OR PSYCHOLOGY EDUCATIONAL OR MUSIC) Timespan = All Years. Databases = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH. Lemmatization = On	543	0	
DOAJ	music AND audiovisual [all fields]	10	0	
	music AND audio-visual [all fields]	7	0	
Total			14 (42.42%)	Bullerjahn & Lehmann (1989), Howard (2009), Min (2001), Peddell (2004), Ryan & Costa-Giomi (2004), Ryan et al. (2006), Schmidt (1976), Siddel-Strebel (2007), Wapnick (2009), Wapnick et al. (1997), Wapnick et al. (1998), Wapnick et al. (2000), Wapnick et al. (2004), Zumpella (1993)

Note: Total number of studies included in systematic review is $n = 33$ (see also Appendixes B and C for a detailed publication list).

published by presenting mostly nonsignificant results (Cooper et al., 2009). This so-called publication or availability bias is an indicator for the existence of unpublished results and investigates how strongly those unpublished studies could influence the result of meta-analysis. To test for the presence of publication biases, we used several approaches: first, the so-called funnel plot (Egger, Smith, Schneider, & Minder, 1997) as an ocular inspection method for the detection of a systematic selection bias of publications (see Figure 1). In case of publication bias, the distribution of results will describe an asymmetrical funnel shape. Figure 1 shows a nearly symmetrical distribution of effect sizes in relation to the precision of estimates as indicated by the standard error. Only one study shows a relatively imprecise estimation (based on a small sample size) and is located along the bottom. However, the funnel plot itself is not robust against the influence of heterogeneity (Irwig, Macaskill, Berry, &

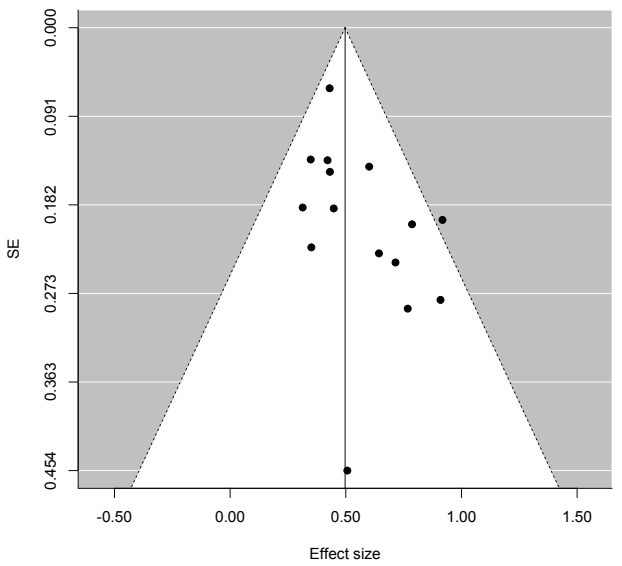

FIGURE 1. Funnel plot of observed studies' effect sizes (Cohen's *d*) against standard error of effect size (*SE*).

Glasziou, 1998) and should be treated with caution. Thus, in a second step, we conducted Egger's linear regression test (Egger et al., 1997). According to Egger et al. (1997), this test examines, by means of regression procedures, if there is an asymmetric distribution in the funnel plot. An asymmetric distribution is confirmed if there is a connection between sample size and effect size estimation. Therefore, every study's effect size had to be transformed to its standard normal deviate ($z_i = \theta_i/s_i$ defined as study's effect size [θ_i] divided by its standard error) that is regressed against its precision ($prec_i = 1/s_i$). In case of no publication bias, the intercept β_0 of this resulting regression should run through the origin ($\beta_0 = 0$). A test of the null hypothesis found no support for any publication bias in our meta-analysis, $t(13) = 2.18$, *ns* (see also Appendix H).

In the next step, we used Orwin's fail-safe *N* method (Orwin, 1983) to identify the number of studies needed to decrease the meta-analysis' effect size to $d = 0.20$ (Appendix I). As a result, additional $n = 22$ studies with an effect size of $d = .00$ would be needed for this decrease. Finally we used the so-called "trim and fill" method (Duval & Tweedie, 2000a, 2000b) for a sensitivity analysis. This algorithm trims off the asymmetric right side of a funnel plot. In this procedure, studies in the right part were replaced by imputed studies on the left side as their missing counterparts (Rothstein et al., 2005). This imputation approach revealed that $n = 4$ studies would be needed to obtain a mathematically "perfect" symmetry (see also Appendix J). Moreover, the resulting imputed mean average effect size only moved 0.06 standard deviations to the left, Cohen's $d_{imp} = 0.45$, 95% CI (0.35, 0.56). Thus, we conclude that there was no significant underlying publication bias in our meta-analysis, which would change the overall observed effect size significantly.

Result

The result from 15 studies on the effect of the visual component on the appreciation of music is summarized in Figure 2 (see Appendix A for statistical details). Based on 1,298 subjects, combined with the effect sizes reported in or recalculated from the 15 studies (Appendix B) our meta-analysis yielded an average weighted effect size of $d = 0.51$ standard deviations for the influence of the visual component on the evaluation of music performance in terms of liking, expressiveness, or overall quality of music … performance (Appendix A). According to Cohen's definition of benchmarks for effect sizes, this medium effect size is "visible to the naked eye" of a careful observer (1988, p. 26). Compared with the IQ distribution, this effect size corresponds to a difference of about 8 IQ points between conditions. The 95% CI (0.42, 0.59) of the average effect size is considerably small and within the positive range, indicating consistent enhancement effects of the visual component. Furthermore, our results revealed that the hypothesis of no effect has to be rejected. Thus, we conclude that our result obtained from a random-effects model is statistically significant at the specified $\alpha = .05$ level ($z = 11.24$, $p < .001$).

Discussion

This meta-analysis revealed a medium effect size in evaluation behavior differences depending on the presentation mode of music performance. Furthermore, considering the small range of the 95% CI around the point estimator, we observed a highly precise estimation of the population effect. We conclude that the visual component is not a marginal phenomenon in music perception, but an important factor in the communication of meaning. This process of cross-modal integration exists for classical as well as pop and rock music (Cook, 2008; Vines et al., 2006). Moreover, the audio-visual performance conveys markers of authenticity (Auslander, 2008), which are essential for the creation of credibility in popular music culture. To put it bluntly, in popular music "seeing is believing" (Auslander, 2008, p. 85). Of course, the result of our meta-analysis is strongly influenced by Hamann's study included in our sample of studies (Hamann, 2003). Her study is based on a large

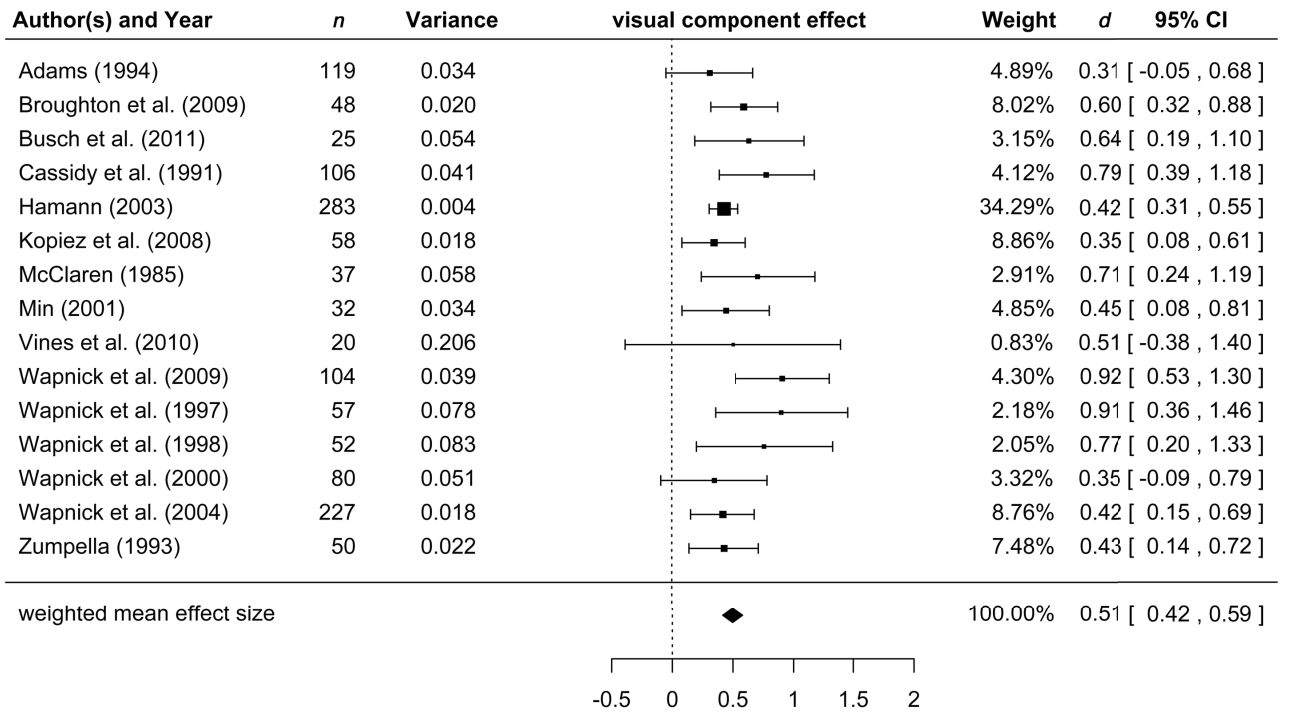

FIGURE 2. Meta-analysis (forest plot) of studies on the effect of the visual component in audio-visual music perception; *n* = number of subjects, *d* = effect size measure, defined as the number of standard deviations by which the audio-visual condition's mean is larger than the audio-only condition's mean.

number of subjects, and thus the observed effect size is characterized by a very high test power. However, some studies are underpowered: These studies are characterized by either nonsignificant results or extensive variance due to small sample sizes and, thus, an underestimation of the observed effect size. As a consequence, these studies are only able to expose vague interval estimations of effect size (e.g., Vines et al., 2011; Wapnick, Mazza, & Darrow, 2000).

Our experience from this meta-analysis shows that many studies are unfortunately characterized by insufficient statistical information. However, complete statistical information is indispensable for later meta-analytical utilization. Thus, to improve this situation, we propose the following criteria for writing the results section of a study: (a) A report of the complete descriptive statistics (means, standard deviations, number of subjects) should be given; (b) ANOVA designs should report all within-cell information including means, standard deviations, and sample sizes as well as sums of squares and mean squares for all effects including nonsignificant effects. The report of *p* values only is insufficient. *F* values, degrees of freedom and exact *p* values should be given up to three decimals, and smaller values should be indicated as $p < .001$; (c) repeated measures or correlated observations designs (MANOVA etc.) should give additional information on all measure-to-measure correlations; (d) information on prospective (a priori) test power ($1-\beta$) for the calculation of sample sizes should be indicated; (e) according to the APA Publication manual (2010, p. 34), information on effect sizes in standardized indicators (e.g., Cohen's *d*, Hedge's *g*, r^2 etc.) are mandatory; (f) materials should be conducive to meta-analysis by including, for example, exact *p* values up to 3 decimals (smaller values should be indicated as $p < .001$), *F* and *t* values, and degrees of freedom for those with nonsignificant results. These aspects can be easily considered at the time of manuscript writing and will contribute to a sustaining use of study results for future researchers. Based on our experience, we know that it is very troublesome and, in some cases, nearly impossible to obtain missing statistical information from authors - especially after more than 10 years from publication.

Up until now, research in multisensory integration of the visual and aural modality has been widely limited to speech perception (McGurk & MacDonald, 1976; Vroomen & de Gelder, 2004; Woods & Recanzone, 2004) or animal behavior (Narins, Grabul, Soma, Gaucher, & Hödl, 2005). Our meta-analysis shows that

multisensory music perception can make a significant contribution to the understanding of the perceptual system. From a psychological perspective, our results also point to the inner structure of the psychic cognitive apparatus: Visual and auditory perceptions are only separated in the periphery, but the inner structure of the psychic apparatus itself is characterized by the complex interaction of senses. Advanced models of intersensory speech perception make clear that the development of analogue approaches for music perception will need statistical values for the determination of intermodal relationships: For example, as Lederman & Klatzky (2004) showed in their conceptual model of intersensory integration of the microstructural surface perception, the question of whether more than one modality (e.g., vision, audio, or touch) improves or impairs perceptual performance needs the indication of weights for each modality. Those weights are crucial for the outcome of the weight generator and the intersensory integrator. As Massaro (2004) shows in his speech-related "fuzzy logical model of perception," the integration process of evaluation and integration combines multisensory input and assumes multiplicative integration for the measurement of the performance output. Thus, we assume that future models of multisensory music perception will benefit from our designation of a verified statistical value.

Against the background of our results, we argue for a higher emphasis on the visual component in the explanation of judgment differences between the two conditions, audio only and audio-visual presentation. The huge evaluation drift depending on the visual component cannot be explained simply as a supportive function in musical communicative settings. As an alternative to simple communication approaches, we argue for the approach of musical persuasion. As already developed in the theory of rhetoric, persuasion is known as the inner core of the rhetoric process (Knape, 2000, p. 33). Furthermore, persuasion is defined as the successful change from the audience's initial mental status into another. In this setting of strategic communication, the musician acts as an orator to gain the audience's favor. This framework could integrate previous findings of audio-visual music performance evaluation research (e.g., Schutz, 2008; Thompson et al., 2005) as well as findings from other domains, such as the social psychology of music performance (for a first approach, see Lehmann & Kopiez, 2011).

A broader implication of our work is that audio-visual signals provide a powerful source of aesthetic communication. Our meta-analysis confirms the importance of visual information, but a converse question is whether musical information makes a significant contribution within largely visual or verbal media, as in television advertisements (see Lalwani, Lwin, & Ling, 2009). Theodor W. Adorno (1968) famously claimed that "Music in television is fuss" (p. 124). Given the appeal of audio-visual communication revealed by our meta-analysis, we question this claim. Future meta-analyses focusing on the role of music and sound in audio-visual contexts would complement our analyses, providing a rich understanding of the relative contributions of sound, music, verbal content, and visual signals for effective multimodal communication.

Author Note

Preparation of this article was supported by PRO*Niedersachsen Grant 76202-23-1/10. We thank Marco Lehmann (University of Hamburg, Germany), Eckart Altenmüller (Hanover University of Music, Drama and Media, Germany) and three anonymous reviewers for helpful comments on a previous version of the manuscript. Supporting online material for the study is available from the website http://musicweb.hmtm-hannover.de/meta-analysis.

Correspondence concerning this article should be addressed to Reinhard Kopiez, Hanover University of Music, Drama and Media, Emmichplatz 1, 30175 Hanover, Germany. E-MAIL: reinhard.kopiez@hmtm-hannover.de

References

Adorno, T. W. (1968). Musik im Fernsehen ist Brimborium [Music in television is fuss]. *Der Spiegel, 9*, 116–124.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Auslander, P. (2008). *Liveness: Performance in a mediatized culture* (2nd ed.). London, UK: Routledge.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823.

Bergeron, V., & Lopes, D. M. (2009). Hearing and seeing musical expression. *Philosophy and Phenomenological Research, 78*, 1–15.

Bermingham, G. A. (2000). Effects of performers' external characteristics on performance evaluations. *Update: Applications of Research in Music Education, 18*, 3–7.

Borenstein, M. (2009). *Introduction to meta-analysis.* Chichester, NH: Wiley.

Borenstein, M. (2010). *Comprehensive meta-analysis* (Version 2.0) [Computer software]. Englewood, NJ: Biostat.

Burger, E. (1986). *Franz Liszt: Eine Lebenschronik in Bildern und Dokumenten* [Franz Liszt: A chronicle in pictures and documents]. München: Liszt Verlag.

Chabris, C. F. (1999). Prelude or requiem for the "Mozart effect"? *Nature, 400*, 826–827.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243–253.

Cook, N. (2008). Beyond the notes. *Nature, 453*, 1186–1187.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles, CA: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. London, UK: SAGE Publications.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*, 103–113.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–99.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634.

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.

Finnäs, L. (2001). Presenting music live, audio-visually or aurally - Does it affect listeners' experiences differently? *British Journal of Music Education, 18*, 55–78.

Frith, S. (1996). *Performing rites: Evaluating popular music.* Oxford, UK: Oxford University Press.

Gabrielsson, A. (2003). Music performance research at the millenium. *Psychology of Music, 31*, 221–272.

Glass, V. G. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5(10)*, 3–8.

Hamann, K. L. (2003). Identification of expressiveness in small ensemble performances by middle school students. *Bulletin of the Council for Research in Music Education, 155*, 24–32.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*, 203–217.

Hetland, L. (2000). Listening to music enhances spatial-temporal reasoning: Evidence for the "Mozart Effect." *Journal of Aesthetic Education, 34*(3/4), 105–148.

Higgins, J. P. T., & Green, S. (2009). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley-Blackwell.

Humphrey, S. E. (2011). What does a great meta-analysis look like? *Organizational Psychology Review, 1*, 99–103.

Irwig, L., Macaskill, P., Berry, G., & Glasziou, P. (1998). Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased. *British Medical Journal, 316*, 470–471.

Juslin, P. N. (2005). From mimesis to catharsis: Expression, perception, and induction of emotion in music. In D. Miell, R. MacDonald, & D. J. Hargreaves (Eds.), *Musical communication* (pp. 85-115). Oxford, UK: Oxford Universty Press.

Kämpfe, J., Sedlmeier, P., & Renkewitz, F. (2010). The impact of background music on adult listeners: A meta-analysis. *Psychology of Music, 39,* 424–448.

Knape, J. (2000). *Was ist Rhetorik?* [What is rhetoric?]. Stuttgart: Reclam.

Lalwani, A. K., Lwin, M. O., & Ling, P. B. (2009). Does audiovisual congruency in advertisements increase persuasion? The role of cultural music and products. *Journal of Global Marketing, 22*, 139–153.

Lederman, S. J., & Klatzky, R. L. (2004). Multisensory texture perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 107–122). Cambridge, MA: MIT Press.

Lehmann, M., & Kopiez, R. (2011). Der Einfluss der Bühnenshow auf die Bewertung der Performanz von Rockgitarristen [The influence of the stage show on the evaluation of rock guitar performance]. In R. F. Nohr & H. Schwaab (Eds.), *Metal Matters: Heavy Metal als Kultur und Welt* (pp. 195–206). Münster: Lit Verlag.

Massaro, D. W. (2004). From multisensory integration to taling heads and language learning. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education, 10*, 12–24.

Narins, P., Grabul, D. S., Soma, K. K., Gaucher, P., & Hödl, W. (2005). Cross-modal integration in a dart-poison frog. *PNAS, 102*, 2425–2429.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational and Behavioral Statistics, 8*, 157–159.

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect - Shmozart effect: A meta-analysis. *Intelligence, 38*, 314–323.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: John Wiley & Sons.

Schumann, R. (1985). *Gesammelte Schriften über Musik und Musiker* [Collected writings on music and musicians] (Vol. 3 & 4). Wiesbaden: Breitkopf & Härtel. (Original work published 1854)

Schutz, M. (2008). Seeing music? What musicians need to know about vision. *Empirical Musicology Review, 3*, 83–108.

Sedlmeier, P. (Ed.). (2009). Beyond the significance test ritual. *Zeitschrift für Psychologie/Journal of Psychology, 217*(1), 1–5.

Seifert, T. L. (1991). Determing effect sizes in various experimental designs. *Educational and Psychological Measurement, 51*, 341–347.

Spreckelmeyer, K. N., Kutas, M., Urbach, T. P., Altenmüller, E., & Münte, T. F. (2006). Combined perception of emotion in pictures and musical sounds. *Brain Research, 1070*, 160–170.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica, 156*, 203–227.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessement Device (Study DIAD). *Psychological Methods, 13*, 130–149.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2011). Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance. *Cognition, 118*, 157–170.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101*, 80–113.

Vroomen, J. & de Gelder, B. (2004). Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 141–150). Cambridge, MA: MIT Press.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performance. *Journal of Research in Music Education, 48*, 323–336.

Woods, T. M., & Recanzone, G. H. (2004). Cross-modal interactions evidenced by the ventriloquism effect in humans and monkeys. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 35–48). Cambridge, MA: MIT Press.

APPENDIX A: Random-effects analysis of weighted mean effect size.

	Statistics for each study						Sample size	
			95% CI					
Study	*d*	*SE*	*LL*	*UL*	*z*	*p*	Treatment	Control
Adams (1994)	0.31	0.18	-0.05	0.68	1.70	.09	59	60
Broughton & Stevens (2009) [rm]	0.60	0.14	0.32	0.88	4.22	.00	48	48
Busch & Wöllner (2011) [rm]	0.64	0.23	0.19	1.10	2.78	.01	25	25
Cassidy & Sims (1991)	0.79	0.20	0.39	1.18	3.90	.00	54	52
Hamann (2003)[rm]	0.43	0.06	0.31	0.55	6.92	.00	283	283
Kopiez & Lehmann (2008)[rm]	0.35	0.14	0.08	0.61	2.58	.01	58	58
McClaren (1985)[rm]	0.71	0.24	0.24	1.19	2.97	.00	37	37
Min (2001)[rm]	0.45	0.19	0.09	0.81	2.42	.02	32	32
Vines et al. (2010)	0.51	0.45	-0.38	1.40	1.12	.27	10	10
Wapnick et al. (2009)	0.92	0.20	0.53	1.30	4.66	.00	60	54
Wapnick et al. (1997)	0.91	0.28	0.36	1.46	3.26	.00	31	26
Wapnick et al. (1998)	0.77	0.29	0.20	1.33	2.66	.00	28	24
Wapnick et al. (2000)	0.35	0.23	-0.09	0.79	1.56	.12	40	40
Wapnick et al. (2004)	0.42	0.14	0.16	0.69	3.10	.00	132	95
Zumpella (1993)[rm]	0.43	0.15	0.14	0.72	2.92	.00	50	50
Total	0.51	0.04	0.42	0.59	11.24	.00		

Note: [rm] = repeated measures design; *d* = standardized difference in means (Cohen's *d*); *SE* = standard error of *d*; 95% CI = confidence interval of *d*; summary = weighted average effect size (random-effects analysis).

APPENDIX B: Studies included in meta-analysis.

Study	Outcome
Adams, B. L. (1994). *The effect of visual/aural conditions on the emotional response to music* (Doctoral Dissertation, Florida State University, Florida, USA). Available from ProQuest Dissertations and Theses database. (UMI No. 9434127)	liking
Broughton, M., & Stevens, C. (2009). Music, movement and marimba: An investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychology of Music, 37*, 137–153.	expressiveness
Busch, V., & Wöllner, C. (2011, September). *Geht es um die Musik? Bewertungen beim Eurovision Song Contest unter der Lupe.*[Does anyone evaluate the music? A closer look at evaluations of the European Song Contest]. Paper presented at the Jahrestagung der Deutschen Gesellschaft für Musikpsychologie: Musik und Gesundheit, Osnabrück, Germany.	liking
Cassidy, J. W., & Sims, W. L. (1991). Effects of special education labels on peers' and adults' evaluations of a handicapped youth choir. *Journal of Research in Music Education, 39*, 23–34.	overall quality
Hamann, K. L. (2003). Identification of expressiveness in small ensemble performances by middle school students. *Bulletin of the Council for Research in Music Education, 155,* 24–32.	overall quality
Kopiez, R., & Lehmann, M. (2008, August). *The influence of the stage show on the evaluation of rock guitar performance.* Paper presented at the 10th International Conference on Music Perception and Cognition (ICMPC 10), Sapporo, Japan.	liking
McClaren, C. A. (1985). *The influence of visual attributes of solo marimbists on perceived qualitative response of listeners* (Doctoral dissertation, The University of Oklahoma, Oklahoma). Available from ProQuest Dissertations and Thesis database. (UMI No. 8524079)	overall quality
Min. P. E. (2001). The effects of visual information on the reliability of evaluation of large instrumental musical ensemble. *Dissertation Abstracts International: Section A. The Humanities and Social Sciences Collection. 62*, 3328.	overall quality
Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2010). Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance. *Cognition, 118*, 157–170.	overall quality
Wapnick, J., Campbell, L., Siddell-Strebel, J., & Darrow, A.-A. (2009). Effects of non-musical attributes and excerpt duration on ratings of high-level piano performances. *Musicae Scientiae, 13*, 35–54.	overall quality
Wapnick, J., Darrow, A. A., Kovacs. J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education, 45*, 470–479.	overall quality
Wapnick, J., Mazza. J. K., & Darrow. A.-A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education, 46*, 510–521.	overall quality
Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness. Stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education, 48*, 323–335.	overall quality
Wapnick, J., Ryan, C., & Lacaille. N. (2004). Effects on selected variables on musicians' ratings of high-level piano performances. *International Journal of Music Education, 22*, 7–20.	overall quality
Zumpella, T. J. (1993). Adjudicated differences in musical performances of high school clarinet students: Audio performances versus audio-visual performances. *Dissertation Abstracts International: Section A. The Humanities and Social Sciences Collection. 55*, 238.	overall quality

APPENDIX C: **Studies excluded from meta-analysis.**

Study	Reason for exclusion
Bullerjahn, C., & Lehmann, A. C. (1989). "Videotraining für Sänger" – zur audiovisuellen Rezeption von Jazz - und Klassikgesang im Fernsehen [Video training for singers – Perception of jazz and classical singing performances on TV]. In K.-E. Behne. G. Kleinen & H. de la Motte-Haber (Eds.). *Musikpsychologie. Jahrbuch der Deutschen Gesellschaft für Musikpsychologie* (Vol. 6, pp. 61–86). Wilhelmshaven: Florian Noetzel Verlag.	no audio-only presentation as control condition
Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*, 103–113.	statistical reporting did not meet our criteria
Geringer, J. M., Cassidy. J. W., & Byo. J. L. (1997). Nonmusic majors' cognitive and affective responses to performance and programmatic music videos. *Journal of Research in Music Education, 45*, 221–233.	statistical reporting did not meet our criteria
Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae Scientiae, 12*, 273–290.	no audio-only presentation as control condition
Howard, S. A. (2009). *The effect of selected nonmusical factors on adjudicators' ratings of high school solo vocal performances* (Doctoral Dissertation, University of Missouri). Available from ProQuest Dissertations and Theses database. (UMI No. 3361571)	statistical reporting did not meet our criteria
Huang, J., & Krumhansl, C. L. (2011). What does seeing the performer add? It depends on musical style, amount of stage behavior, and audience expertise. *Musicae Scientiae*, *15*, 343–364.	statistical reporting did not meet our criteria
Lucas, K. V., & Teachout, D. J. (1998). Identifying expressiveness in small ensemble performances. *Contributions to Music Education, 25*, 60–73.	statistical reporting did not meet our criteria
Lychner, J. A. (2008). A comparison of non-musicians' and musicians' aesthetic response to music experienced with and without music. *International Journal of Music Education, 26*, 21-32.	statistical reporting did not meet our criteria
Madsen, K. (2009). Effect of aural and visual presentation modes on Argentine and US musicians' evaluations of conducting and choral performance. *International Journal of Music Education, 27*, 48–58.	statistical reporting did not meet our criteria
Peddell, L. T. (2004). *Influence of conductor behavior on listeners' perception of expressiveness* (Doctoral Dissertation, University of Minnesota). Available from ProQuest Dissertations and Thesis database. (UMI No. 3137189)	continuous data
Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education, 52*, 141–154.	statistical reporting did not meet our criteria
Ryan, C., Wapnick. J., Lacaille. N., & Darrow, A.-A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music, 34*, 559–572.	same dataset as used in Wapnick et al. (2004)
Schmidt, H.-C. (1976). Auditive und audiovisuelle musikalische Wahrnehmung im experimentellen Vergleich. Fernsehdidaktische Überlegungen für die Sekundarstufe I und II. [Aural and audio-visual music perception with the use of TV. An experimental comparison.] In R. Stephan (Ed.), *Schulfach Musik* (pp. 79–105). Mainz: Schott.	non-parametric statistical procedure
Siddell-Strebel, J. (2007). *The effects of non-musical components on the ratings of performance quality* (Doctoral Dissertation, McGill University, Canada). Available from ProQuest Dissertations and Thesis database. (UMI No. NR32324)	statistical reporting did not meet our criteria
Tan, J. (1999, November). *The effect of modes of presentation on the evaluation of marching band by musicians and nonmusicians.* Paper presented at the Joint AARE - NZARE. Melbourne.	statistical reporting did not meet our criteria
Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education, 53*, 162–176.	statistical reporting did not meet our criteria
Williamon, A. (1999). The value of performing from memory. *Psychology of Music, 27*, 84–95.	no audio-only presentation as control condition
Zembower, C. M. (2000). *The effect of video and audio recordings of concert band performances on adjudicator evaluations* (Doctoral Dissertation, University of Sothern Mississippi, USA). Available from ProQuest Dissertations and Thesis database. (UMI No. 3000266)	statistical reporting did not meet our criteria

APPENDIX D: Research design descriptors.

	Research design descriptors				
	Unit of assignment to conditions	Type of assignment to conditions	Overall confidence of groups at pretest	Outcome	Study design
Adams (1994)	individuals	random	high (strong inference)	liking	independent group
Broughton & Stevens (2009)	individuals		high (strong inference)	expressiveness	repeated measures
Busch & Wöllner (2011)	classroom, facility		high (strong inference)	liking	repeated measures
Cassidy & Sims (1991)	individuals	random	high (strong inference)	overall quality	independent group
Hamann (2003)	classroom, facility		moderate (weak inference)	overall quality	repeated measures
Kopiez & Lehmann (2008)	individuals		high (strong inference)	liking	repeated measures
McClaren (1985)	individuals		moderate (weak inference)	overall quality	repeated measures
Min (2001)	individuals		high (strong inference)	overall quality	repeated measures
Vines et al. (2010)	individuals	random	high (strong inference)	overall quality	independent group
Wapnick et al. (2009)	individuals	cannot tell	moderate (weak inference)	overall quality	independent group
Wapnick et al. (1997)	individuals	random	high (strong inference)	overall quality	independent group
Wapnick et al. (1998)	cannot tell	random	low (guess)	overall quality	independent group
Wapnick et al. (2000)	individuals	random	low (guess)	overall quality	independent group
Wapnick et al. (2004)	classroom, facility	random	low (guess)	overall quality	independent group
Zumpella (1993)	individuals		high(strong inference)	overall quality	repeated measures

APPENDIX E: Data reported in studies for effect size calculation/estimation.

	Effect size data						
	Treatment		Control		Sig. report		Confidence rating in effect size computation
	M	*SD*	*M*	*SD*	*t*	*F*	
Adams (1994)	4.42	0.59	4.23	0.62			no estimation
Broughton & Stevens (2009)*	5.63	0.79	5.16	0.78	4.58		no estimation
Busch & Wöllner (2011) *	4.52	0.96	3.94	0.85	3.05		no estimation
Cassidy & Sims (1991)	2.76	0.91	2.12	0.70	3.67		no estimation
Hamann (2003)*	14.38	2.82	13.19	2.86	7.24		no estimation
Kopiez & Lehmann (2008)*	3.79	0.95	3.56	0.82	2.65		no estimation
McClaren (1985)$^{+}$							moderate estimation
Min (2001)^{re}					2.53		moderate estimation
Vines et al. (2010)*	3.70	0.86	3.25	0.91			no estimation
Wapnick et al. (2009)						23.95	moderate estimation
Wapnick et al. (1997)					3.42		highly estimated
Wapnick et al. (1998)					2.76		highly estimated
Wapnick et al. (2000)					1.57		highly estimated
Wapnick et al. (2004)					3.13		highly estimated
Zumpella (1993)					3.05		highly estimated

Note: Blank cells = not reported; * = author contacted for data; $^{+}$ = effect size estimation for ANOVA designs; ^{re} = re-analysis of published raw data.

APPENDIX F: **Statistical values for fixed- and random-effects models.**

	Raw study data			Fixed-effects sums			Random-effects sums		
Study	*n*	*d*	v_i	*w*	*wd*	Wd^2	w^2	*w**	*w*d*
Adams (1994)	119	0.31	0.034	29.39	9.22	2.89	863.56	26.78	8.40
Broughton & Stevens (2009)	48	0.60	0.020	49.22	29.59	17.79	2422.44	42.31	25.44
Busch & Wöllner (2011)	25	0.64	0.054	18.66	12.01	7.72	348.22	17.57	11.31
Cassidy & Sims (1991)	106	0.79	0.041	24.59	19.34	15.21	604.68	22.74	17.88
Hamann (2003)	283	0.43	0.004	259.04	111.41	47.91	67103.27	139.37	59.94
Kopiez & Lehmann (2008)	58	0.35	0.018	54.68	19.05	6.64	2990.13	46.29	16.13
McClaren (1985)	37	0.71	0.058	17.24	12.33	8.81	297.27	16.31	11.66
Min (2001)	32	0.45	0.034	29.08	13.03	5.84	845.77	26.53	11.88
Vines et al. (2010)	20	0.51	0.206	4.85	2.45	1.24	23.47	4.77	2.42
Wapnick et al. (2009)	104	0.92	0.039	25.72	23.61	21.67	661.48	23.70	21.76
Wapnick et al. (1997)	57	0.91	0.078	12.83	11.66	10.61	164.47	12.30	11.19
Wapnick et al. (1998)	52	0.77	0.083	12.04	9.25	7.10	144.99	11.58	8.89
Wapnick et al. (2000)	80	0.35	0.051	19.70	6.92	2.43	387.96	18.49	6.49
Wapnick et al. (2004)	227	0.42	0.018	54.08	22.77	9.59	2924.15	45.86	19.31
Zumpella (1993)	50	0.43	0.022	45.75	19.73	8.51	2092.57	39.72	17.13
Total	1298			656.85	322.35	173.96	81874.42	494.32	249.82

Note: Number of subjects was corrected for repeated measures. For differentiation between weights of either fixed- or random-effects model, weights of the random-effects model are marked by an asterisk.

APPENDIX G: **Statistics of fixed- and random-effects models.**

	Effect size						Test of null (2-tailed)		Heterogeneity				τ^2	
					95% CI									
Model	*n*	*d*	*SE*	υ	*LL*	*UL*	*z*	*p*	*Q*	*df*	*p*	I^2	τ^2	*SE*
Fixed	15	0.49	0.04	0.002	0.41	0.57	12.58	.00	15.76	14	.33	11.19		
Random	15	0.51	0.05	0.002	0.42	0.59	11.24	.00					0.03	0.01

Note: n = number of studies; υ = variance of effect size; *Q* = test for heterogeneity of effect size parameters; I^2 = proportion of the observed variance reflecting real differences in effect size; τ^2 = between-studies variance.

APPENDIX H: **Egger's linear regression method.**

Intercept				Test of null (2-tailed)		
		95% CI				
β_0	*SE*	*LL*	*UL*	*t*	*df*	*p*
1.13	0.52	0.01	2.25	2.18	13	.05

Note: For further details see Egger et al. (1997)

APPENDIX I: **Orwin's fail-safe *N* for revealing publication bias.**

Parameter	
Std. diff. in means in observed studies	0.49
Criterion for a 'trivial' std. diff. in means	0.20
Mean std. diff. in means in missing studies	0.00
Number of missing studies needed to decrease std. diff. in means to *d* = 0.20	22.00

Note: Std. diff. in means = standard difference in means (Cohen's *d*). The method is based on the assumption of a fixed-effect model.

APPENDIX J: **Sensitivity analysis using "trim and fill"method (Duval & Tweedie, 2000a; b).**

	Effect size			
			95% CI	
Model	*n*	*d*	*LL*	*UL*
Fixed	19	0.44	0.37	0.52
Random	19	0.45	0.35	0.56

Note: "Trim and fill"-method imputed additional four studies.