Memory & Cognition
https://doi.org/10.3758/s13421-020-01060-2

Cognitive and academic benefits of music training with children: A multilevel meta-analysis

Giovanni Sala[1] · Fernand Gobet[2]

Abstract
Music training has repeatedly been claimed to positively impact children's cognitive skills and academic achievement (literacy and mathematics). This claim relies on the assumption that engaging in intellectually demanding activities fosters particular domain-general cognitive skills, or even general intelligence. The present meta-analytic review (N = 6,984, k = 254, m = 54) shows that this belief is incorrect. Once the quality of study design is controlled for, the overall effect of music training programs is null ($\overline{g} \approx 0$) and highly consistent across studies ($\tau^2 \approx 0$). Results of Bayesian analyses employing distributional assumptions (informative priors) derived from previous research in cognitive training corroborate these conclusions. Small statistically significant overall effects are obtained only in those studies implementing no random allocation of participants and employing non-active controls ($\overline{g} \approx 0.200$, $p < .001$). Interestingly, music training is ineffective regardless of the type of outcome measure (e.g., verbal, non-verbal, speed-related, etc.), participants' age, and duration of training. Furthermore, we note that, beyond meta-analysis of experimental studies, a considerable amount of cross-sectional evidence indicates that engagement in music has no impact on people's non-music cognitive skills or academic achievement. We conclude that researchers' optimism about the benefits of music training is empirically unjustified and stems from misinterpretation of the empirical data and, possibly, confirmation bias.

Keywords Academic achievement · Cognitive ability · Cognitive training · Music · Transfer

Introduction

It has been claimed that music fosters children's cognitive skills and academic achievement. Learning to play the violin or the piano, to recognize pitches, and to keep the beat are often presented as effective cognitive enhancement tools (e.g., Jaušovec & Pahor, 2017). However, the idea that practicing cognitively demanding tasks may lead to domain-general cognitive enhancement is in stark contrast with empirical evidence in cognitive science and educational psychology. In fact, while human cognition has been shown to be malleable to training, transfer of skills appears to be limited to the training domain and, at best, other similar domains.

It is customary to distinguish between two broad categories of transfer: near transfer and far transfer (Barnett & Ceci, 2002). Whereas near transfer – i.e., the transfer of skills within the same domain – is sometimes observed, far transfer – i.e., the transfer of skills across two distant domains – is rare or possibly inexistent (Melby-Lervåg, Redick, & Hulme, 2016, Sala et al., 2019a). Moreover, when it does occur, transfer of skills is often limited to the degree to which the two domains (source and target) share contents. For example, even transfer of skills within subspecialties of the same discipline seems to be limited. In fact, performance significantly worsens when experts engage in certain subspecialties of their field of expertise. For example, chess masters who are asked to recall or find the best move in positions coming from chess openings that do not fall into their repertoire exhibit a drastic (about 1 SD) reduction in performance (Bilalić, McLeod, & Gobet, 2009). This so-called *curse of specificity* has been recently defined as one of the fundamental particles in the standard model of human cognition (Sala & Gobet, 2019).

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13421-020-01060-2) contains supplementary material, which is available to authorized users.

✉ Fernand Gobet
F.Gobet@lse.ac.uk

[1] Institute for Comprehensive Medical Science (ICMS), Fujita Health University, Toyoake, Aichi, Japan

[2] Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London WC2A 2AE, UK

Published online: 29 July 2020

Researchers involved in cognitive training do not deny that between-domain, or even within-domain, transfer is hard to trigger. Nonetheless, they claim that it is possible to induce far transfer by engaging in domain-specific cognitively demanding activities that boost domain-general cognitive skills; those skills, in turn, are supposed to generalize across many different domains (e.g., academic proficiency; Strobach & Karbach, 2016). At a neural level, this generalization is thought to be enabled by the activation of shared brain structures that are common to the practiced activity (e.g., music) and other core cognitive skills (e.g., fluid intelligence, working memory, and language; Moreno et al., 2011). In other words, domain-general cognitive enhancement and far transfer are believed to be by-products of domain-specific training (Taatgen, 2016).

With respect to music, three main hypotheses have been formulated to explain why playing it should lead to broad cognitive benefits. To begin with, music might directly impact on general intelligence rather than on some particular cognitive skills (Schellenberg, 2004). This idea is consistent with the vast amount of correlational evidence showing that musicians tend to outperform non-musicians in a variety of cognitive tests. Examples include memory (Sala & Gobet, 2017a; Talamini, Altoè, Carretti, & Grassi, 2017), fluid and general intelligence (Ruthsatz, Detterman, Griscom, & Cirullo, 2008; Schellenberg, 2006), attention (Saarikivi, Putkinen, Tervaniemi, & Huotilainen, 2016), and phonological processing (Forgeard et al., 2008). The same pattern of results occurs in academic skills. In fact, music skills appear to be related to better reading abilities (Anvari, Trainor, Woodside, & Levy, 2002), and music engagement is a predictor of overall academic achievement (Wetter, Koerner, & Schwaninger, 2009).

Another possible link connecting music engagement and cognitive enhancement might be working memory (WM). Multimodal cognitively demanding activities are thought to strengthen WM capacity (Diamond & Ling, 2019; Morrison & Chein, 2011), which, in turn, enhances fluid intelligence and learning (Jaeggi et al., 2008). Music training is one such activity (Saarikivi, Huotilainen, Tervaniemi, & Putkinen, 2019). Simply put, the putative broad benefits of music training would stem from a boost in domain-general WM capacity rather than general intelligence.

Finally, music training might positively impact on one's sound perception and, consequently, phonological processing and even reading skills (Patel, 2011; Tierney & Kraus, 2013). This hypothesis is upheld by the fact that numerous brain structures and neural patterns are shared by music skills and language (for a review, see Jäncke, 2009). Interestingly, improved reading skills may also facilitate the acquisition of new skills and therefore enhance people's IQ performance (Ritchie & Tucker-Drob, 2018). This further mechanism would again be consistent with the overall idea that music training conveys multiple cognitive and academic benefits.

Experimental evidence

The theories just described imply that music training *causes* cognitive enhancement and improvement in academic performance. However, correlational evidence gathered in natural groups is not sufficient to establish a causal link. In the last few decades, dozens of experimental trials have been carried out to examine a potential causal link between music training and improved cognitive/academic performance.

Researchers in this field have reached inconsistent conclusions. While most of them have expressed optimism about the benefits of music training (e.g., Barbaroux, Dittinger, & Besson, 2019; Nan et al., 2018; Tierney, Krizman, & Kraus, 2015), others have found this enthusiasm unjustified (e.g., Kempert et al., 2016; Rickard, Bambrick, & Gill, 2012). Like in many other fields in the social sciences, meta-analyses have been carried out to resolve such controversies. The only comprehensive meta-analytic review performed so far about the benefits of music training is that by Sala and Gobet (2017b). This meta-analysis – which includes 38 studies, 118 effect sizes, and 3,085 participants – found an overall effect of $\overline{d}$ = 0.16. It also highlighted that the impact of music training on cognitive skills and academic performance was a function of the quality of the study's experimental design. Specifically, the magnitude of the music-induced effects was significantly smaller (around zero) in those studies implementing active controls and random allocation of the participants to groups.

Two meta-analyses examined a subset of studies (Cooper, 2019; Gordon, Fehd, & McCandliss, 2015), and drew somewhat more positive implications for the cognitive and educational benefits of music teaching. Gordon et al. (2015) reviewed 12 studies (n = 901) assessing the effects of music training on language-related skills. The overall effect was small but significant ($\overline{d}$ = 0.20). Analogously, Cooper (2019) analysed 21 studies (n = 1,767) and found an overall effect size of $\overline{g}$ = 0.28 across several measures of cognitive ability (measures related to academic achievement were not included because they were considered too different from cognitive ability). Interestingly, the effect was maintained in studies employing active controls ($\overline{g}$ = 0.21).

The present meta-analysis

Despite the less than encouraging evidence, dozens of new experimental investigations have been carried out in recent years, including the two largest randomized control trials (RCTs) in this field (Aleman et al., 2017; Haywood et al., 2015). Once again, the claims about the effectiveness of music training have been inconsistent across studies (e.g., James et al., 2019; Lukács & Honbolygó, 2019; Nan et al., 2018). We thus ran a meta-analysis including both old and new

experimental studies to establish (a) which claims are justified, (b) what are the sources of heterogeneity across studies, and (c) which of the theories predicting that music training enhances cognitive and academic skills are corroborated/refuted.

Beyond being relatively dated, the previous meta-analyses suffer from several technical limitations. First, no multilevel modeling was employed. Multilevel modeling is necessary to adjust standard errors when a certain degree of statistical dependence is present in the data (i.e., effect sizes nested in studies).[1] Also, some of the effect sizes were incorrectly calculated because of a mistake in the reporting of the results in one of the primary studies (Rickard et al., 2012; personal communication). Both issues probably inflated the amount of between-study true heterogeneity, which tended to bias meta-analytic model estimates. In addition, the presence of a non-negligible amount of unexplained true heterogeneity (as in both Sala & Gobet, 2017b, and Cooper, 2019) makes the overall effect sizes hard to interpret because the sources of between-study variability remain hidden. Finally, no thorough sensitivity analysis was performed (e.g., outlier analysis and multiple publication bias analysis). In brief, such suboptimal modeling choices produce biased estimates. The present meta-analytic review aims to correct these problems and to update the findings of the music-training literature. The current meta-analysis also carries out Bayesian analyses that compare the support for the null and alternative hypotheses, and relies on a larger number of studies (19 new studies) and therefore a larger number of participants (an increase from about 3,000 to about 7,000, compared to Sala & Gobet, 2017b). Since the number of participants and effect sizes are more than double, the current meta-analysis has a much higher power than the 2017 meta-analysis.

Method

Literature search

A systematic search strategy was implemented (Appelbaum et al., 2018). Using the following Boolean string ("music" OR "musical") AND ("training" OR "instruction" OR "education" OR "intervention"), we searched through ERIC, PsycINFO, and ProQuest Dissertation & Theses databases to find studies that reported music training programs. We retrieved 3,044 records.

[1] Here, the adjective "multilevel" broadly refers to any technique that allows the researcher to correctly model multivariate data (i.e., effect sizes) nested within studies. We do not intend to imply that the specific modeling method used in the present meta-analysis is superior to other methods. Rather, we simply highlight the necessity of adopting multilevel/multivariate techniques in order to produce accurate results.

Inclusion criteria

Five inclusion criteria were applied:

1) The study was experimental in nature and implemented a cognitively demanding music-training program (e.g., learning to play instruments, Kodály method, etc.). No correlational or ex-post facto studies were included.
2) The study included at least one control group that isolated the variable of interest (i.e., music training).
3) The study included non-music-related cognitive tests or academic outcomes.
4) The study included participants aged between 3 and 16 years with no previous formal music experience or clinical condition.
5) The study reported sufficient data to calculate the effect sizes. Alternatively, the author(s) had to provide the necessary data.

We searched for eligible articles through 1 December 2019. When the data reported in the study were insufficient to calculate the effect sizes or important details about the study design were unclear, we contacted the corresponding authors by email (n = 11). We received three positive replies. We found 54 studies, conducted from 1986 to 2019, that met the inclusion criteria (reported in Appendix A in the Supplemental Online Materials). Nineteen of these studies had never been included in any previous meta-analysis. The studies included 254 effect sizes and a total of 6,984 participants. Thus, compared to the previous most comprehensive meta-analysis in the field (i.e., Sala & Gobet, 2017b), the number of both effect sizes and participants was more than doubled. The studies originally evaluated for inclusion but eventually excluded are reported in Appendix B in the Supplemental Online Materials. The procedure is described in Fig. 1.

Moderators

We assessed six moderators based on the previous meta-analyses in the literature:

1) Baseline difference (continuous variable): The standardized mean difference between the experimental and control groups at pre-test. This moderator was added to evaluate the amount of true heterogeneity accounted for by pre-post-test regression to the mean. It thus aimed at ruling out potential confounding effects of this statistical artifact.
2) Randomization (dichotomous variable): Whether the children were randomly allocated to the groups.
3) Type of controls (active or non-active; dichotomous variable): Whether the music training group was compared to another novel activity (e.g., dancing); no-contact groups and business-as-usual groups were treated as "non-

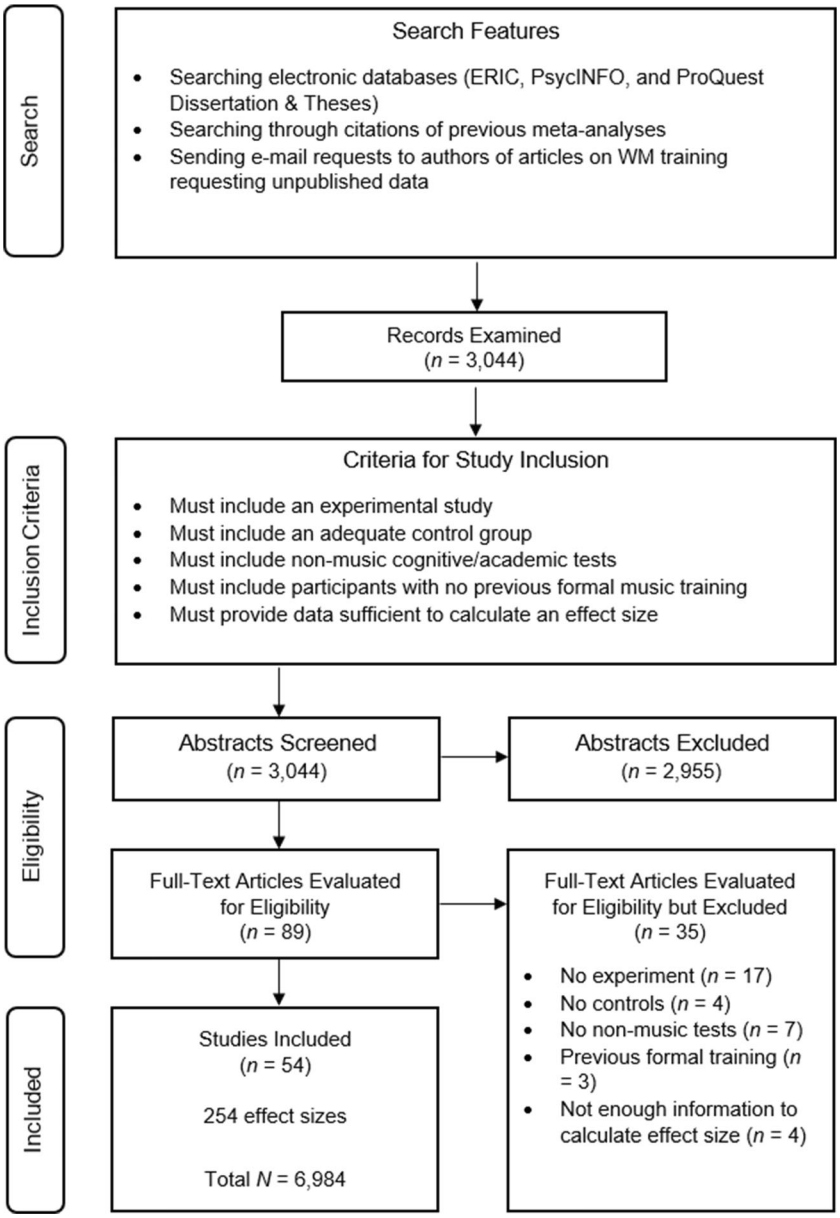

Fig. 1 Flow diagram of the search strategy

active." This moderator thus controlled for potential placebo effects.

4) Age (continuous variable): The mean age of the study's participants. A few studies did not report the participants' mean age. In these cases, the participants' mean age was obtained from the median or the school grade.
5) Outcome measure: The effect sizes were grouped into four broad groups based on the Cattell-Horn-Carroll taxonomy (McGrew, 2009): *non-verbal ability* (e.g., fluid reasoning [Gf], mathematical skills [Gq], and spatial skills [Gv]); *verbal ability* (e.g., vocabulary and reading skills [Gc], phonological processing [Grw]); *memory* (e.g., short-term/working-memory tasks [Gsm]); and *speed* (e.g., processing speed [Gs] and inhibition tasks [Gt]). The inter-rater agreement was $\kappa = 1$. We also examined this moderator without grouping these eight categories into the four groups. Finally, since some primary studies employed academic tests (e.g., Haywood et al., 2015), we examined whether the effect sizes related to cognitive skills were greater than those related to academic achievement (as suggested by Cooper, 2019). The latter category included all those effect sizes that were obtained from academic tests of literacy and mathematics (a subset of the Gc and Gq groups). All the other effect sizes fell into the cognitive group.
6) Duration of training: The total number of hours, weeks, and sessions of the training program in each study. These three variables were tested separately because they were

collinear (i.e., they measured the same construct with three different metrics).

Effect size calculation

The effect sizes were calculated for each eligible outcome measure in the primary studies. Hedges's *g*s – an adjusted standardized mean difference – were calculated with the following formula:

$$g = d \times \left(1 - \frac{3}{(4 \times N) - 9}\right) \quad (1)$$

with

$$d = \frac{(M_{e\_post} - M_{e\_pre}) - (M_{c\_post} - M_{c\_pre})}{SD_{pooled\_pre}} \quad (2)$$

where M_{e_post} and M_{e_pre} are the mean of the experimental group at post-test and pre-test, respectively, M_{c_post} and M_{c_pre} are the mean of the control group at post-test and pre-test, respectively, SD_{pooled_pre} is the pooled pre-test SDs in the experimental group and the control group, and N is the total sample size.

The sampling error variances were calculated with the following formula:

$$Var_g = \left(\frac{N_e - 1}{N_e - 3} \times \left(\frac{2 \times (1 - r)}{r_{xx}} + \frac{d_e^2}{2} \times \frac{N_e}{N_e - 1}\right) \times \frac{1}{N_e} + \frac{N_c - 1}{N_c - 3} \times \left(\frac{2 \times (1 - r)}{r_{xx}} + \frac{d_c^2}{2} \times \frac{N_c}{N_c - 1}\right) \times \frac{1}{N_c}\right) \times \left(1 - \frac{3}{(4 \times N) - 9}\right)^2 \quad (3)$$

where r_{xx} is the test-retest reliability of the test, N_e and N_c are the sample sizes of the experimental group and the control group, respectively, d_e and d_c are the within-group standardized mean differences of the experimental group and the control group, respectively. Finally, r is the pre-post-test correlation (Schmidt & Hunter, 2015; pp. 343–355). The pre-post-test correlations and test-retest coefficients were rarely provided in the primary studies. Therefore, we assumed the reliability coefficient (r_{xx}) to be equal to the pre-post-test correlation (i.e., no treatment by subject interaction was postulated; Schmidt & Hunter, 2015; pp. 350–351), and we imposed the pre-post-test correlation to be $r_{xx} = r = .600$.

When the study implemented an only-post-test design (i.e., no pre-test assessment) we used the following formulas for effect size and sampling error variance, respectively:

$$g = \frac{M_{e\_post} - M_{c\_post}}{SD_{pooled\_pre}} \times \left(1 - \frac{3}{(4 \times N) - 9}\right) \quad (4)$$

$$Var_g = \frac{N - 1}{N - 3} \times \frac{4}{N} \times \left(1 + \frac{d^2}{8}\right) \times \left(1 - \frac{3}{(4 \times N) - 9}\right)^2 \quad (5)$$

Finally, in a few cases, *t*- and *F*-values were used to calculate *d* (for the details, see the Supplemental Online Materials).

Modeling approach

Robust variance estimation (RVE) with correlational weights was employed to perform the intercept and meta-regression models (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016). RVE has been designed to model nested effect sizes (i.e., extracted from the same study). Two indexes were used to report the models' between-cluster true (i.e., not due to random error) heterogeneity: τ^2, which indicates the absolute amount of true heterogeneity; and I^2, which indicates the percentage of true heterogeneity. In addition, we manipulated the within-study effect-size correlation (ρ) assumed by the RVE models to test the sensitivity of the results to this parameter. We performed these analyses with the Robumeta R package (Fisher, Tipton, & Zhipeng, 2017).

Publication bias

We examined publication bias with two methods: Duval and Tweedie's (2000) trim-and-fill analysis and Vevea and Woods' (2005) selection models. The trim-and-fill method estimates whether some smaller-than-average effect sizes have been suppressed from the literature and calculates an adjusted overall effect size and standard error. This analysis was conducted after averaging the statistically dependent effects using Cheung and Chan's (2014) approach. We employed the *L0* and *R0* estimators designed by Duval and Tweedie (2000). Vevea and Woods' (2005) selection models estimate publication bias and calculate an adjusted overall effect size (but no standard error) by assigning to *p*-value ranges different weights. In other words, the method assumes that the probability of an effect not to be suppressed is a function of its *p*-value. As recommended by Pustejovsky and Rodgers (2019), the weights used in the publication bias analyses were not a function of the effect sizes (for more details, see Appendices C and D in the Supplemental Online Materials). We performed these analyses with the Metafor R package (Viechtbauer, 2010).

True heterogeneity and sensitivity analysis

Explaining between-study true heterogeneity is one of the main goals of meta-analysis. While small to null true heterogeneity indicates that between-study differences are merely an artifact of random error (Schmidt, 2010), large amounts of true heterogeneity suggest that more than one true effect is present in the data. Moreover, true heterogeneity reduces the statistical power of meta-analytic models, tends to artificially inflate overall effect sizes in asymmetric distributions, and sometimes produces biased publication-bias adjusted estimates (Cheung & Chan, 2014; Henmi & Copas, 2010; Schmidt & Hunter, 2015; Stanley, 2017).

Investigating the sources of true heterogeneity is thus essential to make the results more interpretable and accurate. Therefore, beyond running meta-regression analysis, we performed a two-step sensitivity analysis. First, we excluded three studies that, probably due to lack of random allocation or small sample sizes, reported unusually high between-group differences (≈ 1 SD) in the participants' baseline IQ (Patscheke, Degé, & Schwarzer, 2019; Roden, Kreutz, & Bongard, 2012; Roden, Grube, Bongard, & Kreutz, 2014). That is, these three studies were included in the main analysis but removed from the sensitivity analysis. Such large baseline differences make any findings hard to interpret and may introduce noise in the data. Second, we ran Viechtbauer and Cheung's (2010) influential case analysis. This method evaluates whether some effect sizes exerted an unusually strong influence on the model's parameters such as the amount of between-study true heterogeneity (τ^2). Those effect sizes that inflated true heterogeneity were excluded.

Bayesian analysis

A vast quantity of data regarding cognitive-training programs has been collected in the last 15 years. For example, Sala et al.'s (2019a) second-order meta-analysis estimates that more than 20,000 participants have undergone cognitive-training programs such as music training, videogame training, and WM training. This previous evidence can be employed to establish a set of distributional assumptions (informative priors) in the Bayesian framework.

The distribution of the effect sizes was assumed to be normal. Based on Sala et al.'s (2019a) second-order meta-analysis, we expected the mean effect size to be null (prior $\overline{g} = 0$) in models including active control groups and slightly positive (prior $\overline{g} = 0.150$) in models including passive controls groups. The prior for the standard deviation was the same in all the models ($SD_g = 0.500$). The true heterogeneity parameter (τ) was assumed to have a half-Cauchy distribution (centered on 0 and scale $\gamma = 10$) in all the models. No further prior was used for other moderators.

We thus estimated the Bayes factors (*BF*s) for two sets of competing hypotheses for $\overline{g}$ and τ. First, we compared the alternative hypothesis H1: $\overline{g} \neq 0$ with the null hypothesis H0: $\overline{g} = 0$. Second, we compared the alternative hypothesis H1: $\tau > 0$ with the null hypothesis H0: $\tau = 0$. *BF*s > 1 indicated support for H1, while *BF*s < 1 indicated support for H0. In line with common guidelines, H1 was considered as substantially supported only if BF > 3 (i.e., H1 three times more likely to be true than H0; e.g., Dougherty, Hamovitz, & Tidwell, 2016). Analogously, H0 was substantially supported only if *BF* < 0.333 (i.e., H0 three times more likely to be true than H1). Since the priors were conditional to the type of controls employed by the primary study (as indicated by Sala et al., 2019a), these analyses were carried out after running moderator analysis. The analyses were carried out with the bayesmeta R package (Röver, 2017).

Results

Descriptive statistics

The mean age of the samples was 6.45 years. The median age was 5.90, the first and third quartiles were 5.03 and 7.85, and the mean age range was 3.50–11.59. The mean Baseline difference was -0.038, the median was 0, the first and third quartiles were -0.210 and 0.141, and the range was -1.058–0.844. The mean duration of training was 53.37 h (range 2.00–507.00, median 30.00), 29.29 weeks (range 3.00–117.00, median 26.00), and 53.43 sessions (range 6.00–195.00, median 30.00). The descriptive statistics of the categorical moderators are reported in Table 1.

Table 1 Number of studies and effect sizes sorted by categorical moderators

Moderator	No. of studies	No. of effect sizes
Randomization		
Non-random	33	139
Random	23	115
Control group		
Non-active	41	144
Active	23	110
Outcome measures		
Memory	19	57
Verbal	33	89
Non-verbal	27	69
Speed	13	39

Main analyses

The overall effect size of the RVE intercept model was $\overline{g}$ = 0.184, *SE* = 0.041, 95% confidence interval (CI) [0.101; 0.268], *m* = 54, *k* = 254, *df* = 38.36, *p* < .001, τ^2 = 0.041, I^2 = 43.16%. Different values of within-study effect-size correlation (ρ) did not significantly affect the results ($\overline{g}$ range 0.184– 0.185, τ^2 = 0.041). The random-effect (RE) model (with Cheung and Chan's correction) yielded very similar estimates: $\overline{g}$ = 0.176, *SE* = 0.037, *p* < .001, τ^2 = 0.033. Overall, the results showed a small and moderately heterogeneous overall effect of music training on cognitive and academic outcomes. The results were not affected by modeling choices (i.e., ρ values and procedure for modeling nested data).

Baseline difference and Type of controls were the only two statistically significant moderators (*p* = .031 and *p* = .035, respectively) and accounted for part of the true heterogeneity (τ^2 = 0.038, I^2 = 34.87%). Age was not significant (*p* = .403), neither was Allocation (*p* = .518). No significant differences were found across the four broad groups of outcome measures (all *p*s ≥ .624; Holm's correction for multiple comparisons), nor across the more fine-grained categorization (eight levels, all *p*s ≥ .362), and there was no difference between cognitive skills and measures of academic achievement (*p* = .981). Duration of training was not significant either (*p* = .266, *p* = .952, and *p* = .662 for hours, weeks, and sessions, respectively).

Type of controls

Since Type of controls was statistically significant, we performed the analyses on the two sub-samples separately. In those studies that implemented non-active controls, the results showed a small and moderately heterogeneous overall effect of music training on cognitive and academic outcomes. The overall effect size was $\overline{g}$ = 0.228, SE = 0.045, 95% CI [0.137; 0.320], m = 41, k = 144, df = 30.1, p < .001, τ^2 = 0.042, I^2 = 43.11%. Different values of within-study effect-size correlation (ρ) did not affect the results ($\overline{g}$ = 0.228, τ^2 = 0.042). The RE model provided similar results, $\overline{g}$ = 0.201, SE = 0.041, p < .001, τ^2 = 0.023. Again, the results were not affected by modeling choices. Also, some evidence of a small publication bias was found. The trim-and-fill retrieved no missing study with the L0 estimator. Five missing studies were retrieved with the R0 estimator, and the adjusted estimate was $\overline{g}$ = 0.170, 95% CI [0.064; 0.276]. Vevea and Woods' (2005) selection model calculated a similar estimate ($\overline{g}$ = 0.119). Finally, the Bayes factors confirmed these findings. BF_g was greater than 730,000, indicating that $\overline{g}$ was far more likely to be non-null (H1: $\overline{g} \neq 0$) than null (H0: $\overline{g} = 0$). Regarding the model's true heterogeneity, BF_τ was greater than 5,000, again indicating that τ was far more likely to be positive than null.

In those studies that implemented active controls, the results showed a near-zero and slightly heterogeneous overall effect of music training on cognitive and academic outcomes. The overall effect size was $\overline{g}$ = 0.056, *SE* = 0.058, 95% CI [-0.069; 0.182], *m* = 23, *k* = 110, *df* = 12.6, *p* = .350, τ^2 = 0.025, I^2 = 23.10%. Different values of within-study effect-size correlation (ρ) did not significantly affect the results ($\overline{g}$ range 0.054–0.057, τ^2 range 0.023–0.025). The results were robust to the type of modeling approach employed. In fact, the RE model provided similar results, $\overline{g}$ = 0.090, *SE* = 0.060, *p* = .136, τ^2 = 0.032. Some evidence of a small publication bias was found, suggesting that the unbiased overall effect size is essentially null. No missing study was retrieved with the *L0* estimator, whereas the *R0* estimator identified four missing studies and the adjusted estimate was $\overline{g}$ = -0.020, 95% CI [-0.183; 0.142]. The selection model estimate was $\overline{g}$ = 0.039. The Bayes factors were BF_g = 1.231 and BF_τ = 0.044. These results showed that – as indicated by the publication-bias-corrected estimates – $\overline{g}$ was not convincingly more likely to be non-null than null (BF_g < 3), and that τ was approximately 23 times more likely to be null than positive. The latter finding confirms that the low observed true heterogeneity (τ^2 = 0.025, I^2 = 23.10%) is very likely to be spurious.

Sensitivity analyses

This section replicated the analyses after excluding the three studies reporting large baseline IQ differences across the groups and implemented Viechtbauer and Cheung's (2010) influential case analysis to explain the model's residual true heterogeneity (if any). The overall effect size of the RVE intercept model was $\overline{g}$ = 0.166, *SE* = 0.041, 95% CI [0.083; 0.249], *m* = 51, *k* = 235, *df* = 34.9, *p* < .001, τ^2 = 0.036, I^2 = 40.62%. Different values of within-study effect-size correlation (ρ) did not significantly affect the results ($\overline{g}$ range 0.165–0.166, τ^2 range 0.035–0.036). The random-effect (RE) model provided similar estimates: $\overline{g}$ = 0.149, *SE* = 0.035, *p* < .001, τ^2 = 0.024. Baseline difference and Type of controls were again the only two statistically significant moderators (*p* = .017 and *p* = .003, respectively) and accounted for part of the true heterogeneity (τ^2 = 0.029, I^2 = 29.70%). Therefore, the results were pretty much the same as in the main analyses so far.

Non-active controls

When non-active controls were used, the overall effect size was $\overline{g}$ = 0.226, *SE* = 0.045, 95% CI [0.133; 0.319], *m* = 40, *k* = 139, *df* = 29.2, *p* < .001, τ^2 = 0.041, I^2 = 42.96%. Different values of within-study effect-size correlation (ρ) did not significantly affect the results ($\overline{g}$ = 0.226, τ^2 = 0.041). The RE model provided similar results, $\overline{g}$ = 0.200, *SE* = 0.041, *p* < .001, τ^2 = 0.024. Five effect sizes were found to be significantly inflating the true heterogeneity. After excluding these

effect sizes, the overall effect size was $\overline{g}$ = 0.181, *SE* = 0.042, 95% CI [0.093; 0.268], *m* = 39, *k* = 134, *df* = 21.9, *p* < .001, τ^2 = 0.018, I^2 = 24.92%. Similar results were obtained with the RE model, $\overline{g}$ = 0.161, *SE* = 0.037, *p* < .001, τ^2 = 0.013.

Finally, in order to investigate the sources of the unexplained true heterogeneity (τ^2 = 0.018, I^2 = 24.92%), a moderator analysis was run. Randomization was the only statistically significant moderator (*p* = .042) and explained nearly all the true heterogeneity (τ^2 = 0.005, I^2 = 7.61%). Therefore, the observed true between-study heterogeneity in the studies employing non-active controls was accounted for by a few extreme effect sizes and the type of allocation of participants to the groups. For non-randomized studies, the overall effect sizes were $\overline{g}$ = 0.246, *SE* = 0.049, 95% CI [0.140; 0.352], *p* < .001; for randomized studies, the relevant statistics were $\overline{g}$ = 0.064, *SE* = 0.065, 95% CI [-0.116; 0.244], *p* = .381. Thus, when random allocation was employed, the overall effect was near-zero.

Publication bias analysis: Studies without randomization With the studies that did not implement any randomization of participants' allocation, the trim-and-fill analysis retrieved two missing studies with the *L0* estimator (adjusted estimates $\overline{g}$ = 0.211, 95% CI [0.095; 0.328]). Three missing studies were retrieved with the *R0* estimator (adjusted estimates $\overline{g}$ = 0.189, 95% CI [0.068; 0.310]). Vevea and Woods' (2005) selection model calculated a more conservative estimate ($\overline{g}$ = 0.126). Thus, a small amount of publication bias was still detected. The Bayes factors were BF_g = 217.840 ($\overline{g}$ far more likely to be non-null than null) and BF_τ = 0.021 (τ nearly 50 times more likely to be null than positive). While confirming that the overall effect size of music training in non-randomized samples and passive controls is positive (yet small), these results showed that no between-study true heterogeneity was present in the data.

Publication bias analysis: Studies with randomization Regarding the randomized samples, all the publication bias analyses estimated a substantially null overall effect. The trim-and-fill analysis retrieved six and ten studies with the *L0* and *R0* estimators, respectively (adjusted estimates $\overline{g}$ = 0.009, 95% CI [-0.095; 0.113] and $\overline{g}$ = -0.034, 95% CI [-0.131; 0.063]). Vevea and Woods' (2005) selection model yielded a similar estimate ($\overline{g}$ = -0.002). The Bayes factors were BF_g = 0.257 and BF_τ = 0.025. Therefore, the Bayes factors provided compelling evidence that both $\overline{g}$ and τ are more likely to be null than non-null (approximatively 4 and 40 times, respectively).

Active controls

Turning our attention to the studies implementing active controls, the overall effect size was $\overline{g}$ = -0.021, *SE* = 0.032, 95% CI [-0.109; 0.068], *m* = 20, *k* = 96, *df* = 4.2, *p* = .558, τ^2 = 0, I^2 = 0%. Different values of within-study effect-size correlation (ρ) did not affect the results ($\overline{g}$ = -0.021, τ^2 = 0). The RE model provided similar results, $\overline{g}$ = -0.010, *SE* = 0.035, *p* = .787, τ^2 = 0. Since this model showed no true heterogeneity and null overall effects, no publication bias analysis was performed. The Bayes factors largely favored the null hypothesis (BF_g = 0.063 and BF_τ = 0.006). The null hypothesis was approximatively 16 times and 180 times more likely than the alternative hypothesis for $\overline{g}$ and τ, respectively. In brief, all the analyses showed that the overall effect in studies implementing active controls is null and homogeneous across studies (i.e., $\overline{g}$ = 0, τ^2 = 0).

Discussion

This meta-analytic review investigated the impact of music training on children's cognitive skills and academic achievement. The overall impact of music training programs on cognitive and academic outcomes is weak and moderately heterogeneous ($\overline{g}$ = 0.184, *SE* = 0.041, τ^2 = 0.041, I^2 = 43.16%). The inspection of true heterogeneity shows that there is an inverse relationship between the studies' design quality and magnitude of the effect sizes. Specifically, those studies using active controls or implementing random assignment report homogeneous null or near-zero effects ($\overline{g}$ = -0.021–0.064, $\tau^2 \leq 0.005$). Conversely, a small overall effect size is observed in those studies employing neither active controls nor random assignment ($\overline{g}$ = 0.246). The results of the Bayesian analyses corroborate the conclusions that the unbiased effect of music training on cognitive and academic skills is null and highly consistent across the studies (i.e., $\overline{g}$ = 0 and τ^2 = 0). No other study features (e.g., age, duration of training, and outcome measure) seem to have any influence on the effect sizes – not even the outcome measures. In particular, contrary to Cooper's (2019) hypothesis, there was no difference between cognitive skills and academic achievement (literacy and mathematics), which means that it is justifiable to pool the two outcomes together, as was done for example in Sala and Gobet (2017b). Altogether, these results indicate that music training fails to produce solid improvements in all the examined cognitive and academic skills equally. Finally, only a low amount of publication bias is observed in the models (about 0.100 standardized mean difference at most), which is in line with the near-zero effect sizes estimated. The results are summarized in Table 2.

These findings confirm and extend the conclusions of the previous comprehensive meta-analysis in the field (Sala & Gobet, 2017b). Along with re-establishing the fundamental role of design quality in affecting the experimental results, the present meta-analysis has succeeded in explaining *all* the observed true heterogeneity. We can thus conclude that these

Table 2 Overall effects in the meta-analytic models

Model (1)	$\overline{g}$.RVE (SE) (2)	Adj.$\overline{g}$ (range) (3)	Heterogeneity (4)	Residual heterogeneity (5)	Bayes factors (6)
Main analyses					
Overall	0.184 (0.041)	–	$\tau^2 = 0.041$, $I^2 = 43.16$	$\tau^2 = 0.038$, $I^2 = 34.87$	–
Non-active	0.228 (0.045)	0.119 – 0.228	$\tau^2 = 0.042$, $I^2 = 43.11$	–	$BF_g > 7.3\times10^5$, $BF_\tau > 5\times10^3$
Active	0.056 (0.058)	-0.020 – 0.056	$\tau^2 = 0.025$, $I^2 = 23.10$	–	$BF_g = 1.231$, $BF_\tau = 0.044$
Sensitivity analyses					
Overall	0.166 (0.041)	–	$\tau^2 = 0.036$, $I^2 = 40.62$	$\tau^2 = 0.029$, $I^2 = 29.70$	–
Non-active	0.226 (0.045)	–	$\tau^2 = 0.041$, $I^2 = 42.96$	$\tau^2 = 0.005$, $I^2 = 7.606$	–
Non-random	0.246 (0.049)	0.126 – 0.211	–	–	$BF_g = 217.840$, $BF_\tau = 0.021$
Random	0.064 (0.065)	-0.034 – 0.009	–	–	$BF_g = 0.257$, $BF_\tau = 0.025$
Active	-0.021 (0.032)	–	$\tau^2 = 0.000$, $I^2 = 0.000$	–	$BF_g = 0.063$, $BF_\tau = 0.006$

Note. (1) The meta-analytic model; (2) the overall RVE effect size (Standard Error); (3) the range of the publication bias adjusted estimates; (4) the amount of true heterogeneity of the model; (5) the true heterogeneity after running meta-regression (and sensitivity analysis); (6) Bayes factors comparing the alternative hypotheses (H1: $\overline{g} \neq 0$; H1: $\tau > 0$) with the null hypotheses (H0: $\overline{g} = 0$; H0: $\tau = 0$)

findings convincingly refute all the theories claiming that music training *causes* improvements in any domain-general cognitive skill or academic achievement (e.g., Moreno et al., 2011; Patel, 2011; Saarikivi et al., 2019; Tierney & Kraus, 2013). In fact, there is no need to postulate any explanatory mechanism in the absence of any genuine effect or between-study variability. In other words, since there is no phenomenon, there is nothing to explain. More broadly, these results establish once again that far transfer – due to the very nature of human cognition – is an extremely rare occurrence (Gobet & Simon, 1996; Sala & Gobet, 2019).

Beyond meta-analytic evidence

It is worth noting that other researchers have reached the same conclusions using different methodologies. To begin with, Mosing, Madison, Pedersen, and Ullén (2016) have investigated the relationship between music training and general intelligence in twins. Notably, music-trained twins do not possess a higher IQ than non-music-trained co-twins. This study thus suggests that engaging in music has no effect on people's IQ. Swaminathan, Schellenberg, and Khalil (2017) show that music aptitude, rather than the amount of music training, predicts fluid intelligence in a sample of adults. This finding upholds the idea that the correlation between intelligence and engagement in music is mediated by innate (as opposed to trained) music skills. Similarly, Swaminathan, Schellenberg, and Venkatesan (2018) demonstrate that the correlation between amount of music training and reading ability in adults disappears when domain-general cognitive skills are controlled for.

These findings corroborate the hypothesis according to which the observed correlation between music training and particular domain-general cognitive/academic skills is a byproduct of previous abilities. Once pre-existing differences in overall cognitive function are ruled out, the correlation disappears (Swaminathan & Schellenberg, 2019). Therefore, there is no reason to support the hypothesis that music training boosts cognition or academic skills. Rather, all the evidence points toward the opposite conclusion, that is, that the impact of music training on cognitive and academic skills is null.

Finally, the failure of music-training regimens to induce any generalized effect is mirrored by findings in other cognitive-training literatures. For instance, WM training does not enhance children's domain-general cognitive skills or academic achievement (Aksayli, Sala, & Gobet, 2019; Melby-Lervåg et al., 2016; Sala & Gobet, 2020). The same applies to action and nonaction videogame training and brain training (Duyck & Op de Beeck, 2019; Kassai, Futo, Demetrovics, & Takacs, 2019; Libertus et al., 2017; Lintern & Boot, 2019; Sala et al., 2019a; Sala, Tatlidil, & Gobet, 2018, 2019b; Simons et al., 2016).

The perception of music training effectiveness is biased

It is our conviction that, while the data show a consistent picture, the narrative that has been built around music training is substantially distorted. For example, Schellenberg (2019) has shown how correlational evidence is often used by scholars to incorrectly infer causal relationships between engagement in music and non-music outcomes. Correlation is notoriously insufficient to establish causal links between variables, which makes Schellenberg's (2019) findings quite concerning. Interestingly, this problem appears to be particularly severe in neuroscientific studies.

The overall interpretation of the results reported in the primary studies is another example of the extent to which authors

sometimes misread the empirical evidence presumably supporting music training. For instance, Barbaroux et al.'s (2019) study does not implement any type of controls, which makes their results uninterpretable. Tierney et al. (2015) report non-significant and inconsistent effects on language-related outcomes between a music training group and an active control group. However, this study is not experimental because the participants were recruited after they had chosen what activity to take part in (i.e., self-selection of the sample). (This is why, incidentally, this study is not included in the present meta-analysis.) Despite representing very little evidence in favor of a causal link between music training and improved cognitive/academic skills, the study has gained a considerable amount of attention in news outlets and among researchers in the field (top 5% in Altmetric). In the same vein, Nan et al. (2018) have found no significant effect of music training on any two music-related measures and no effect at all on the examined non-music outcome measures. (The paper reports a barely significant effect [$p = .044$] in an auditory task that is obtained with an ANOVA performed on the mean pre-post-test gains. This is a well-known incorrect practice that inflates Type I error rates.) Therefore, this study corroborates the idea that the impact of music training on cognitive/academic skills is slim to null. Nonetheless, both the authors and several news outlets provide an over-optimistic, if not utterly incorrect, view of the benefits of music training (e.g., McCarthy, 2018).

By contrast, the two largest randomized controlled trials in the field have been either somewhat ignored (Aleman et al., 2017) or nearly completely overlooked (Haywood et al., 2015) by researchers involved in music training (and news outlets). Both studies report no effect of music training on any cognitive or academic skills. Neither of them makes any overstatement about the benefits of music training on any domain-general cognitive or academic skill. It is thus apparent that if *all* the results are considered and correctly interpreted, the whole music training literature depicts a very consistent mosaic. What is mixed is how the same findings are described by different scholars (Schmidt, 2017).

Conclusions and recommendations for future research

This meta-analysis has examined the experimental evidence regarding the impact of music training on children's non-music cognitive skills and academic achievement. The ineffectiveness of the practice is apparent and highly consistent across studies. Moreover, recent correlational studies have confirmed that music engagement is not associated with domain-general cognitive skills or academic performance.

Two alternative potential avenues involving music activities may be worth some interest. First, music may be beneficial for non-cognitive constructs in children such as prosocial behavior and self-esteem (e.g., Aleman et al., 2017). These possible advantages are not likely to be specific to music, though. In fact, any enticing and empowering activity may improve children's well-being. Second, elements of music instruction (e.g., arithmetical music notation) could be used to facilitate learning in other disciplines such as arithmetic (Azaryahu, Courey, Elkoshi, & Adi-Japha, 2019; Courey, Balogh, Siker, & Paik, 2012; Ribeiro & Santos, 2017). Too few studies have been conducted to reach a definite conclusion. Nonetheless, this approach is undoubtedly more likely to succeed than the music training programs reviewed in this meta-analysis. In fact, while the latter program regimens have tried and failed to reach cognitive enhancement via music training, the former methodology tries to convey domain-specific knowledge by focusing on domain-specific information. This type of near transfer is notoriously much easier to achieve (Gobet, 2016; Gobet & Simon, 1996).

Author note A previous version of this meta-analysis was presented at the 41st Annual Meeting of the Cognitive Science Society in Montréal (July 2019). The article published in the Conference Proceedings reports some pilot analyses on a subset of the studies included in the present version.

Open Practices Statement The data that support the findings of this study are openly available in OSF at https://osf.io/rquye/.

References

Aksayli, N. D., Sala, G., & Gobet, F. (2019). The cognitive and academic benefits of Cogmed: A meta-analysis. *Educational Research Review, 29*, 229-243. https://doi.org/10.1016/j.edurev.2019.04.003

Aleman, X., Duryea, S., Guerra, N. G., McEwan, P. J., Muñoz, R., Stampini, M., & Williamson, A. A. (2017). The effects of musical training on child development: A randomized trial of El Sistema in Venezuela. *Prevention Science, 18*, 865-878. https://doi.org/10.1007/s11121-016-0727-3

Anvari, S. H., Trainor, L. G., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology, 83*, 111-130. https://doi.org/10.1016/S0022-0965(02)00124-8

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for

quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*, 3-25. https://doi.org/10.1037/amp0000191

Azaryahu, L., Courey, S. J., Elkoshi, R., & Adi-Japha, E. (2019). 'MusiMath' and 'Academic Music' - Two music-based intervention programs for fractions learning in fourth grade students. *Developmental Science*. Advanced online publication. https://doi.org/10.1111/desc.12882

Barbaroux, M., Dittinger, E., & Besson, M. (2019). Music training with Démos program positively influences cognitive functions in children from low socio-economic backgrounds. *PLoS ONE, 14*: e0216874. https://doi.org/10.1371/journal.pone.0216874

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128, 612-637. https://doi.org/10.1037/0033-2909.128.4.612

Bilalić, M., McLeod, P., & Gobet, F. (2009). Specialization effect and its influence on memory and problem solving in expert chess players. *Cognitive Science, 33*, 1117-1143. https://doi.org/10.1111/j.1551-6709.2009.01030.x

Cheung, S. F., & Chan, D. K. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavioral Research Methods, 46*, 331-345. https://doi.org/10.3758/s13428-013-0386-2

Cooper, P. K. (2019). It's all in your head: A meta-analysis on the effects of music training on cognitive measures in schoolchildren. *International Journal of Music Education*. Advanced online publication. https://doi.org/10.1177/0255761419881495

Courey, S. J., Balogh, E., Siker, J. R., & Paik, J. (2012). Academic music: Music instruction to engage third-grade students in learning basic fraction concepts. *Educational Studies in Mathematics, 81*, 251-278. https://doi.org/10.1007/s10649-012-9395-9

Diamond, D., & Ling, D. S. (2019). Review of the evidence on, and fundamental questions about, efforts to improve executive functions, including working memory. In J. M. Novick, M. F. Bunting, M. R. Dougherty, & R. W. Engle (Eds.), *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development* (pp. 145-389). https://doi.org/10.1093/oso/9780199974467.003.0008

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2016). Reevaluating the effectiveness of *n*-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review, 23*, 206-316. https://doi.org/10.3758/s13423-015-0865-9

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 276-284. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Duyck, S., & Op de Beeck, H. (2019). An investigation of far and near transfer in a gamified visual learning paradigm. *PLoS ONE 14*: e0227000. https://doi.org/10.1371/journal.pone.0227000

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). Package "robumeta." Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

Forgeard, M., Schlaug, G., Norton, A., Rosam, C., Iyengar, U., & Winner, E. (2008). The relation between music and phonological processing in normal-reading children and children with dyslexia. *Music Perception, 25*, 383-390. https://doi.org/10.1525/mp.2008.25.4.383

Gobet, F. (2016). *Understanding expertise: A multi-disciplinary approach*. London: Palgrave/Macmillan.

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted positions. Implications for the theory of expertise. *Memory & Cognition, 24*, 493-503. https://doi.org/10.3758/BF03200937

Gordon, R. L., Fehd, H. M., & McCandliss, B. D. (2015). Does music training enhance literacy skills? A meta-analysis. *Frontiers in Psychology, 6*:1777. https://doi.org/10.3389/fpsyg.2015.01777

Haywood, S., Griggs, J., Lloyd, C., Morris, S., Kiss, Z., & Skipp, A. (2015). *Creative futures: Act, sing, play. Evaluation report and executive summary*. London: Educational Endowment Foundation.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65. https://doi.org/10.1002/jrsm.5

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine, 29*, 2969-2983. https://doi.org/10.1002/sim.4029

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008) Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*, 6829-6833. https://doi.org/10.1073/pnas.0801268105

James, C. E., Zuber, S., Dupuis-Lorenzon, E., Abdili, L., Gervaise, D., & Kliegel, M. (2019). Formal string instrument training in a class setting enhances cognitive and sensorimotor development of primary school children. *bioRxiv*. https://doi.org/10.1101/829077

Jäncke, L. (2009). The plastic human brain. *Restorative Neurology and Neuroscience, 27*, 521-538. https://doi.org/10.3233/RNN-2009-0519

Jaušovec, N., & Pahor, A. (2017). *Boost your IQ with music*. Retrieved from http://scitechconnect.elsevier.com/boost-your-iq-with-music/

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin, 145*, 165-188. https://doi.org/10.1037/bul0000180

Kempert, S., Götz, R., Blatter, K., Tibken, C., Artelt, C., Schneider, W., & Stanat, P. (2016). Training early literacy related skills: To which degree does a musical training contribute to phonological awareness development? *Frontiers in Psychology, 7*:1803. https://doi.org/10.3389/fpsyg.2016.01803

Libertus, M. E., Liu, A., Pikul, O., Jacques, T., Cardoso-Leite, P., Halberda, J., & Bavelier, D. (2017). The impact of action video game training on mathematical abilities in adults. *AERA Open, 3*, 2332858417740857. https://doi.org/10.1177/2332858417740857

Lintern, G., & Boot, W. (2019). Cognitive training: Transfer beyond the laboratory? *Human Factors The Journal of the Human Factors and Ergonomics Society*. Advance online publication. https://doi.org/10.1177/0018720819879814

Lukács, B., & Honbolygó, F. (2019). Task-dependent mechanisms in the perception of music and speech: Domain-specific transfer effects of elementary school music education. *Journal of Research in Music Education, 67*, 153-170. https://doi.org/10.1177/0022429419836422

McCarthy, A. (2018). *Music lessons can improve language skills*. Retrieved from https://www.sott.net/article/390725-Music-lessons-can-improve-language-skills

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10. https://doi.org/10.1016/j.intell.2008.08.004

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of far-transfer: Evidence from a meta-analytic review. *Perspective on Psychological Science, 11*, 512-534. https://doi.org/10.1177/1745691616635612

Moreno, S., Bialystok, E., Barac, R., Schellenberg, E. G., Cepeda, N. J., & Chau, T. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological Science, 22*, 1425-1433. https://doi.org/10.1177/0956797611416999.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review, 18*, 46-60. https://doi.org/10.3758/s13423-010-0034-0

Mosing, M. A., Madison, G., Pedersen, N. L., & Ullén, F. (2016). Investigating cognitive transfer within the framework of music

practice: Genetic pleiotropy rather than causality. *Developmental Science, 19*, 504-512. https://doi.org/10.1111/desc.12306

Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C. C., Dong, Q., … Desimone, R. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. *PNAS, 115*, E6630-E6639. https://doi.org/10.1073/pnas.1808412115

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology, 2*, 142. https://doi.org/10.3389/fpsyg.2011.00142

Patscheke, H., Degé, F., & Schwarzer, G. (2019). The effects of training in rhythm and pitch on phonological awareness in four- to six-year-old children. *Psychology of Music, 47*, 376-391. https://doi.org/10.1177/0305735618756763

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods, 10*, 57-71. https://doi.org/10.1002/jrsm.1332

Ribeiro, F. S., & Santos, F. H. (2017). Enhancement of numeric cognition in children with low achievement in mathematic after a non-instrumental musical training. *Research in Developmental Disabilities, 62*, 26-39. https://doi.org/10.1016/j.ridd.2016.11.008

Rickard, N. S., Bambrick, C. J., & Gill, A. (2012). Absence of widespread psychosocial and cognitive effects of school-based music instruction in 10-13-year-old students. *International Journal of Music Education, 30*, 57-78. https://doi.org/10.1177/0255761411431399

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*, 1358-1369. https://doi.org/10.1177/0956797618774253

Roden, I., Grube, D., Bongard, S., & Kreutz, G. (2014). Does music training enhance working memory performance? Findings from a quasi-experimental longitudinal study. *Psychology of Music, 42*, 284-298. https://doi.org/10.1177/0305735612471239

Roden, I., Kreutz, G., & Bongard, S. (2012). Effects of a school-based instrumental music program on verbal and visual memory in primary school children: A longitudinal study. *Frontiers in Psychology, 3*: 572. https://doi.org/10.3389/fpsyg.2012.00572

Röver, C. (2017). Bayesian random-effects meta-analysis using the bayesmeta R package. https://arxiv.org/abs/1711.08683

Ruthsatz, J., Detterman, D., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence, 36*, 330-338. https://doi.org/10.1016/j.intell.2007.08.003

Saarikivi, K. A., Huotilainen, M., Tervaniemi, M., & Putkinen, V. (2019). Selectively enhanced development of working memory in musically trained children and adolescents. *Frontiers in Integrative Neuroscience, 13*:62. https://doi.org/10.3389/fnint.2019.00062

Saarikivi, K., Putkinen, V., Tervaniemi, M., & Huotilainen, M. (2016). Cognitive flexibility modulates maturation and music-training-related changes in neural sound discrimination. *European Journal of Neuroscience, 44*, 1815-1825. https://doi.org/10.1111/ejn.13176

Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019a). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology, 5*, 18. https://doi.org/10.1525/collabra.203

Sala, G., & Gobet, F. (2017a). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. *Memory & Cognition, 45*, 183-193. https://doi.org/10.3758/s13421-016-0663-2

Sala, G., & Gobet, F. (2017b). When the music's over. Does music skill transfer to children's and young adolescents' cognitive and academic skills? A meta-analysis. *Educational Research Review, 20*, 55-67. https://doi.org/10.1016/j.edurev.2016.11.005

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*, 9-20. https://doi.org/10.1016/j.tics.2018.10.004

Sala, G., & Gobet, F. (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-019-01681-y

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin, 144*, 111-139. https://doi.org/10.1037/bul0000139

Sala, G., Tatlidil, K. S., & Gobet, F. (2019b). Still no evidence that exergames improve cognitive ability: A commentary on Stanmore et al. (2017). *Neuroscience and Biobehavioral Reviews*. Advanced online publication. https://doi.org/10.1016/j.neubiorev.2019.11.015

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science, 15*, 511-514. https://doi.org/10.1111/j.0956-7976.2004.00711.x

Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology, 98*, 457-468.

Schellenberg, E. G. (2019). Correlation = causation? Music training, psychology, and neuroscience. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000263.

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*, 233-242. https://doi.org/10.1177/1745691610369339

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology, 5*, 32-41. https://doi.org/10.1037/arc0000033

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Newbury Park, CA: Sage.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S.E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17*, 103-186. https://doi.org/10.1177/1529100616661983

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*, 581-591. https://doi.org/10.1177/1948550617693062

Strobach, T., & Karbach, J. (Eds.) (2016). *Cognitive training: An overview of features and applications*. New York: Springer.

Swaminathan, S., & Schellenberg, E. G. (2019). Musical ability, music training, and language ability in childhood. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0000798

Swaminathan, S., Schellenberg, E. G., & Khalil, S. (2017). Revisiting the association between music lessons and intelligence: Training effects or music aptitude? *Intelligence, 62*, 119-124. https://doi.org/10.1016/j.intell.2017.03.005

Swaminathan, S., Schellenberg, E. G., & Venkatesan, K. (2018). Explaining the association between music training and reading in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 992-999. https://doi.org/10.1037/xlm0000493

Taatgen, N. A. (2016). Theoretical models of training and transfer effects. In: Strobach T., Karbach J. (eds) *Cognitive Training* (pp. 19-29). Springer, Cham.

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLoS One 12*: e0186773. https://doi.org/10.1371/journal.pone.0186773

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2*, 85-112. https://doi.org/10.1007/s40865-016-0026-5

Tierney, A., & Kraus, N. (2013). Music training for the development of reading skills. *Progress in Brain Research, 207,* 209-241. https://doi.org/10.1016/B978-0-444-63327-9.00008-4

Tierney, A. T., Krizman, J., & Kraus, N. (2015). Music training alters the course of adolescent auditory development. *PNAS, 112*, 10062-10067. https://doi.org/10.1073/pnas.1505114112

Vevea, J. L. & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*, 428-443. https://doi.org/10.1037/1082-989X.10.4.428

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software, 36*, 1-48. Retrieved from http://brieger.esalq.usp.br/CRAN/web/packages/metafor/vignettes/metafor.pdf

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112-125. https://doi.org/10.1002/jrsm.11

Wetter, O. E., Koerner, F., & Schwaninger, A. (2009). Does musical training improve music performance? *Instructional Science, 37*, 365-374. https://doi.org/10.1007/s11251-008-9052-y