Jahrbuch Musikpsychologie
Yearbook of Music Psychology

PsychOpen GOLD

Research Reports

The Acquisition of Tonal Hierarchies in Western Music During School Years: A Re-Analysis of 40 Years of Research

Der Erwerb tonaler Hierarchien in westlicher Musik während der Schulzeit: Eine Re-Analyse der Ergebnisse aus 40 Jahren Forschung

Hanna Mütze [1], Friedrich Platz [2], Veronika Busch [1]

[1] *Department of Musicology and Music Education, University of Bremen, Bremen, Germany.* **[2]** *Institut für Ästhetisch-Kulturelle Wissenschaft und Praxis, Europa-Universität Flensburg, Flensburg, Germany.*

Jahrbuch Musikpsychologie, 2025, Vol. 33, Article e205, https://doi.org/10.5964/jbdgm.205

Received: 2024-08-26 • **Accepted:** 2025-09-22 • **Published (VoR):** 2025-11-12

Reviewed by: Reinhard Kopiez; Klaus Frieler; Thomas Schäfer.

Corresponding Author: Hanna Mütze, Department of Musicology and Music Education, University of Bremen, Bibliothekstraße 1, 28359 Bremen, Germany. E-mail: muetzeh@uni-bremen.de

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

Understanding the relationships between different pitches as a form of tonality is a key element of listening skills in Western tonal music. Tonal hierarchies (i.e., genre-dependent differing prominence of tones) are reflected in the internal representations of tonal hierarchies (IRTH) in long-term memory. Over the past 40 years, research on how individuals—primarily students aged 6 to 15, as well as adults—acquire IRTH has yielded varied and sometimes contradictory conclusions about the timeline and underlying mechanisms of this process. This review aims to synthesize the evidence and critically examine potential reasons for the heterogeneity in prior findings. To this end, two approaches were applied. First, a Bayesian three-level meta-analysis of 60 effect sizes from 16 studies, reported in 13 articles, revealed a medium difference in IRTH sensitivity between younger and older participants. Second, a model comparison analysis based on cross-sectional data from a single study revealed a non-linear growth dynamic, with a larger increase during adolescence as the best model solution to describe the relationship between sensitivity and age. We also examined the considerable heterogeneity observed within and between studies, particularly how task-specific features of the operationalizations might account for these differences. These findings contribute to the development of theoretical models of music-related skill acquisition and suggest directions for future research.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

Keywords

tonal development, probe tone, Bayesian meta-analysis, musical skill acquisition

Zusammenfassung

Das Verständnis der Beziehungen zwischen verschiedenen Tonhöhen als eine Form der Tonalität ist ein zentraler Bestandteil der Hörfähigkeit in westlich-tonaler Musik. Tonale Hierarchien (d. h. die genreabhängig unterschiedliche Prominenz von Tönen) spiegeln sich in internen Repräsentationen tonaler Hierarchien (IRTH) im Langzeitgedächtnis wider. In den vergangenen 40 Jahren hat die Forschung darüber, wie Teilnehmende – hauptsächlich Schulkinder im Alter von 6 bis 15 Jahren, aber auch Erwachsene – IRTH erwerben, unterschiedliche und mitunter widersprüchliche Schlussfolgerungen hinsichtlich des zeitlichen Ablaufs und der zugrunde liegenden Mechanismen hervorgebracht. In diesem Beitrag sollen die Erkenntnisse zusammengefasst und mögliche Gründe für die Heterogenität der Ergebnisse kritisch untersucht werden. Zu diesem Zweck wurden zwei Ansätze angewandt. Erstens ergab eine Bayesianische Meta-Analyse von 60 Effektgrößen aus 16 Studien, die in 13 Artikeln berichtet wurden, einen Medianunterschied der IRTH-Empfindlichkeit zwischen jüngeren und älteren Kindern im Umfang einer mittleren Effektgröße. Zweitens zeigte eine Modellvergleichsanalyse auf der Grundlage von Querschnittsdaten aus einer einzigen Studie eine nichtlineare Wachstumsdynamik mit einem größeren Anstieg während der Adoleszenz und nicht im Kindes-/Grundschulalter als beste Modelllösung zur Beschreibung der Beziehung zwischen Sensibilität und Alter. Eine Heterogenitätsanalyse zeigte abschließend, dass sich die Verteilung der Effektstärken früherer Studien maßgeblich auf aufgabenspezifische Merkmale im Zuge der Operationalisierungen der Repräsentation tonaler Hierarchien zurückführen und erklären lassen. Diese Ergebnisse tragen zur Entwicklung theoretischer Modelle über den Erwerb musikbezogener Fertigkeiten bei und eröffnen Perspektiven für zukünftige Forschungen.

Schlüsselwörter

tonale Entwicklung, Prüftonverfahren, bayessche Metaanalyse, musikalischer Fertigkeitserwerb

Tonality is reflected by the variations in prominence given to tones on a scale (Piston, 1978; von Helmholtz, 1896) because of differences in their frequency of occurrence, metrical position, and duration (Prince & Schmuckler, 2014; Verosky, 2021). Most Western music considers the concept of tonality as:

> ...one of the main conceptual categories of Western musical thought, [referring] to the orientation of melodies and harmonies towards a referential (or tonic) pitch class. In the broadest possible sense, however, it refers to systematic arrangements of pitch phenomena and relations between them. (Hyer, 2021, p. 1)

Several music psychology models suggest that tonality perception is linked to the varying stabilities perceived in individual tones (Schmuckler, 2016). Krumhansl and Shepard (1979) demonstrate that tonal hierarchy is reflected in individuals' stability ratings of

PsychOpen GOLD

tones, depending on the hierarchical level in the preceding tonal context. This hierarchy in classical/romantic music includes four levels: tonic, tonic triad, other diatonic tones, and non-diatonic tones (Krumhansl & Kessler, 1982). Krumhansl and Keil (1982) conclude that tonal hierarchy is stored in long-term memory as internal representations of tonal hierarchies (IRTH). In this paper, the ability to perceive the different stabilities of tones is referred to as sensitivity to IRTH.

Based on the above, our analyses focus on the IRTH acquisition processes (i.e., how IRTH evolves with increasing age). The scientific debate about the timing and mechanisms of IRTH development is ongoing (Patel, 2021), with narrative summaries (e.g., Corrigall & Schellenberg, 2016; Gembris, 2017a) and theoretical models offering varied and sometimes contradictory conclusions. Earlier theories, such as Brehmer's concept of "tonal giftedness" (German: "Begabung"), suggest that this ability is fully developed by age 7 (Kühn, n.d., cited in Brehmer, 1925, p. 172). Gordon (2012) argues that musical aptitude and audiation show significant development before stabilizing around age 9, with limited improvement thereafter. This extended developmental trajectory contrasts with Brehmer's earlier views.

Gardner's (1973) theory includes two stages of artistic development concluding around age 7, whereas Swanwick and Tillman (1986) propose four stages extending into adolescence. Hargreaves (1996) supports the latter view, emphasizing the importance of ages 8 to 15 for mastering musical rules. Conversely, Serafine (1988) suggests a more restricted sensitive period for tonal learning from ages 8 to 10. In stark contrast, recent research highlights statistical learning as a key factor in lifelong IRTH development (Jonaitis & Saffran, 2009; Vuvan, 2013), with evidence showing that this process begins as early as infancy (Saffran et al., 1999). Thus, whether IRTH has already developed fully by one's school years or continues to develop during that time remains an open question.

Regarding primary studies, Krumhansl and Keil (1982) identify a clear age-related progression in IRTH: first- and second-graders prefer diatonic tones over non-diatonic tones, third- and fourth-graders prefer the tonic triad over other diatonic tones, and fifth-graders as well as adults prefer the tonic overall. Maier-Karius and Schwarzer (2011), Paananen (2007, 2009), and Schwarzer et al. (1993) replicate this trajectory; however, Matsunaga et al. (2020), Schellenberg et al. (2005), Speer and Meeks (1985), and Wilson and Wales (1995) find either earlier or later IRTH acquisition.

The failure to replicate previous findings may be related to several factors, such as participant characteristics or methodological differences. One factor might be formal musical training, defined as systematic instruction in music theory or practice (Hanna-Pladdy & MacKay, 2011). Although some studies suggest that formal training enhances IRTH acquisition (Corrigall et al., 2022; Mandikal Vasuki et al., 2016), others find no such effect, primarily attributing acquisition to implicit learning processes (Cui, 2019; Müllensiefen et al., 2014). Moreover, the influence of formal training may depend on factors such as age, ongoing training, and varying types of operationalization (Asztalos

PsychOpen GOLD

& Csapo, 2017; Müllensiefen et al., 2022; Zhang et al., 2020). Cultural background (Matsunaga et al., 2020), gender (Lin, 2023), and socioeconomic status (Miles et al., 2016) may also modulate IRTH acquisition.

Studies may also suffer from a lack of statistical power, which increases the likelihood of failing to detect a true population effect (Ellis, 2010). Finally, the specific method used to measure IRTH may affect the results. In contrast to the predictions of earlier theories on the development of tonal hierarchies, which rely on explicit measures with (partially) time-independent response formats, several studies using implicit measures report detecting tonal knowledge even in early childhood, indicating an earlier onset of this ability than previously suggested (Jentschke et al., 2005; Politimou et al., 2021; Trehub et al., 1999). Corrigall et al. (2022) find that task characteristics influence IRTH acquisition in school-aged children, with significant differences only emerging in explicit tasks.

In summary, despite 40 years of vibrant research, we are faced with a highly heterogeneous body of literature, making it difficult to draw clear evidence on the shape of IRTH skill acquisition. This complexity pertains not only to identifying the general developmental trajectory but also to distinguishing between specific phases of skill acquisition, such as the phase of initial skill acquisition and the potential onset of a saturation effect—a plateau in skill improvement in relation to the conditions (e.g., formal instruction) under which differences in the shapes and effects of skill acquisition occur.

Aims and Objectives

We conducted a critical review of four decades of research on the development of IRTH sensitivity. Despite a sizeable volume of studies, the findings remain heterogeneous and partially contradictory, particularly regarding developmental trajectories and the influence of musical training. Building on Gordon's (2012) theory of music learning, which posits that core aspects of tonal sensitivity are largely developed by age 9, we formulated four exploratory research questions. By examining these questions, this study aimed to (1) assess the magnitude of age-related changes in IRTH sensitivity, (2) examine whether sensitivity increases significantly beyond age 9, (3) investigate the role of musical training, and (4) explore the effects of different operationalizations of IRTH measurement.

Research Questions

Research Question 1 (RQ1). How substantial is the average age-related development of IRTH sensitivity?
Research Question 2 (RQ2). Does IRTH sensitivity continue to increase beyond the age of 9, or does it level off, as predicted by Gordon's learning theory?
Research Question 3 (RQ3). Is IRTH sensitivity significantly higher in musically trained individuals than in those without musical training?

PsychOpen GOLD

Research Question 4 (RQ4). Does the implicit measurement of IRTH sensitivity via response time yield lower estimates than more explicit operationalizations?

Method

Procedure

We conducted a systematic and comprehensive literature search (What Works Clearinghouse, 2020) between January and April 2022 to identify eligible studies. The study selection followed a predefined process outlined in the review protocol (see Supplementary Material S1), which was not published before the review was conducted. In the final sample, we included only primary research studies of the preliminary literature corpus that used a hypothesis-testing approach, followed either a cross-sectional or longitudinal design, were published between 1982[1] and April 2022, investigated a sample of healthy elementary and secondary school students, and defined the measurement of IRTH sensitivity as a dependent variable based on either a listening task (probe tone paradigm, response time measures, and goodness-of-fit ratings in syntax-violation paradigms) or a creative production task (harmonization of a given melody, tonal composition, and improvisation to a given chord sequence). These operationalizations were further specified as described below.

First, the probe tone paradigm (Krumhansl & Shepard, 1979) involves presenting one or two of the 12 chromatic tones (the "probe") after establishing the tonal context (e.g., an ascending major scale), followed by a brief silence. Participants rate how well the probe tones fit the tonal context. This process continues until all 12 chromatic tones are rated. The perceived stability of each tone is measured based on participants' goodness-of-fit ratings. The correlation of these ratings with the prototypical tonal hierarchy (see above) defines participants' sensitivity to IRTH. Although no standardized protocols or psychometric criteria[2] have been established, the probe tone paradigm has been consistently replicated across various contexts, demonstrating its robustness (Morgan et al., 2019; Sauvé et al., 2021).

Second, some studies use goodness-of-fit ratings in a syntax-violation paradigm (e.g., Corrigall et al., 2022). Participants judge how well the final tone of short melodies fits within the preceding harmonic context with varying degrees of tonal congruence (e.g., a tonic note such as "C" in C major versus a non-diatonic note such as "C#"). A

1) We included only articles published after 1982 following Krumhansl and Kessler's (1982) pioneering work in this field.

2) The lack of these psychometric criteria also applies to all of the measurement methods (exception: Wilson & Wales, 1995).

PsychOpen GOLD

greater difference in ratings between congruent and incongruent endings indicates IRTH acquisition, with higher ratings for congruent endings reflecting higher IRTH sensitivity.

Third, Schellenberg et al. (2005) and Corrigall et al. (2022) focus on response times (as introduced by Janata & Reisberg, 1988). Rather than explicitly evaluating the tonal congruence of melody endings, participants rate other musical features, such as the timbre of the final tone, while hearing a priming sequence with either a tonally congruent (expected) or incongruent (unexpected) final tone. Faster and more accurate responses are expected when the target tone or chord is more tonally congruent with the preceding context, indicating stronger IRTH sensitivity.

Fourth, other methods involve creative production tasks, such as composing (Wilson & Wales, 1995), improvising while listening to a pre-recorded tonal chord sequence (Paananen, 2003), or harmonizing a melody (Paananen, 2009). The outcome variables are defined by the degree of tonal fit and timing of the tones or chords used by the participants. Wilson and Wales (1995, p. 102) report a substantial to almost perfect inter-rater agreement (cf., Landis & Koch, 1977) for expert assessments of compositions ($.69 \leq \kappa \leq .92$).

In the next step, three independent and trained coders (including the first author) coded the included studies according to a previously developed protocol (see Supplemental Material S2). After calculating the initial inter-rater reliability, which showed almost perfect agreement ($\kappa = .92$), coding discrepancies were resolved by consensus. The studies vary in whether and how they report participants' formal musical training; therefore, we calculated the percentage of those with training for each age group. Sufficient data are available in 11 studies, while other variables are coded as not applicable ("n/a").

Data Analysis

Data were analyzed using two approaches. First, we conducted a Bayesian three-level meta-analysis to provide a quantitative summary of the systematic review, offer an overview, identify notable studies, and investigate differences in IRTH measurements. Second, we conducted a model comparison analysis of the cross-sectional data of a single study (Krumhansl & Keil, 1982) investigating the growth trajectory of IRTH acquisition.

In the meta-analytical approach, effect sizes were estimated from primary studies using the metafor package (Viechtbauer, 2010) in R (R Core Team, 2022) and, if needed, transformed into Cohen's *d* following Borenstein and Hedges (2019, pp. 214–234). The procedure for effect size calculations, raw data, and a markdown script for the Bayesian three-level meta-analysis are shown in Supplemental Materials S3, S4, and S5.

In the subsequent analysis step, the effect sizes were aggregated and weighted using a Bayesian three-level meta-analysis with the R package brms (Bürkner, 2017), following the procedure of Harrer et al. (2021). Unlike a one-level fixed effect model that only attributes variance to sampling error, the two-level random effects model separates var-

PsychOpen GOLD

iance into sampling errors[3] and between-study heterogeneity (τ_1^2). This model captures inherent differences between studies, such as measurement methods.

Our model further includes a third level of variance (τ_2^2), reflecting the nested structure of effect sizes within studies and accounting for the variance within single studies, such as different age groups. This multilevel approach (for a formal model description, see Supplemental Material S6) better reflects the nested structure of our data and avoids the statistical issues of traditional univariate meta-analysis, which can misestimate heterogeneity and increase the risk of false positives (Harrer et al., 2021; Hedges, 2019). We compared models with differing levels of complexity by computing marginal likelihoods using the bridge sampling method implemented in the brms package.

We chose Bayesian meta-analysis for two key reasons. First, it allows for the explicit modeling of heterogeneity uncertainty (τ^2), which can yield more stable estimates in small-sample contexts—especially when applying weakly or moderately informative priors (Harrer et al., 2021). Notably, although Bayesian methods are not inherently robust to small samples or poor study quality, they provide a principled framework for incorporating prior knowledge and quantifying uncertainty. Second, prior information can improve the posterior estimation of key parameters, such as μ (intercept) and τ (standard deviations; Röver, 2020).

To address the context-dependence of choosing the prior parameter settings (Gelman et al., 2015), we initially used two separate and weakly informative zero-centered priors for μ and τ, as recommended by Williams et al. (2018) and Harrer et al. (2021, Setting Prior Distributions, para. 5):

- Prior 1: $\mu \sim \text{Normal}(0,\ 1)$
- Prior 2: $\tau \sim \text{Halfcauchy}(0,\ 0.5)$

Subsequently, we performed sensitivity analyses to estimate the impact of prior parameter choices on the posterior density distributions of the parameter estimates. Therefore, following Röver (2020) and Turner and Higgins (2019), we used two alternative weakly informative priors for both μ and τ, while τ was modeled as a standard deviation parameter constrained to be positive, implying half-distributions for the respective priors:

- Prior 3: $\mu,\ \tau \sim \text{Normal}(0,\ 0.5)$
- Prior 4: $\mu,\ \tau \sim t(3,\ 0,\ 2.5)$

We did not conduct a meta-regression using age as a moderator because of the limited number of studies, heterogeneity in age reporting, and insufficient information on age distribution in several samples.

3) As a consequence of Turner and Higgins (2019, p. 302), Bayesian models do not estimate the sampling error, as it is assumed to be known based on the sample size.

PsychOpen GOLD

We examined the magnitude of age-related development in IRTH sensitivity (RQ1) by fitting a Bayesian three-level meta-analytic model using the brms package in R. Small effects ($d \approx 0.2$) are often considered the minimum threshold for practical relevance in developmental research (e.g., Gignac & Szodorai, 2016), and the average difference in IRTH sensitivity between older and younger participants meets or exceeds the benchmark for a small effect. Insufficient information reported in the primary studies prevented us from conducting a meta-regression with musical training as a moderator (RQ2). Therefore, the impact of musical training on IRTH development remains an open research question.

To address RQ3, we estimated the impact of operationalization on effect size using a Bayesian three-level meta-regression in brms. The model included operationalization as a moderator, accounted for the nesting of effect sizes within studies, and incorporated their standard errors. Random effects were specified at the study and within-study levels.

We examined RQ4 using Krumhansl and Keil (1982), who report notable within-study heterogeneity, to compare learning-theory-informed models using data from. The dataset comprised four age groups assessed across five probe tone conditions. Based on age-specific difference scores, we fitted linear (linear, quadratic, cubic) and non-linear (sigmoidal, logistic, mixed, saturated) functions to model the development of tone judgment stability. Analyses were conducted using the R packages lme4 (Bates et al., 2015) and nlme (Pinheiro & Bates, 2024; see S7 for data and S8 for code).

Results

Our search process yielded h = 3,635 hits. We excluded articles based on a range of criteria (see Figure 1 for details). Ultimately, we included i = 13 articles reporting j = 16 studies with y = 60 effect sizes for N = 1,287 participants.

Table 1 presents the descriptive statistics for the effect sizes of the included studies. All studies were cross-sectional, with no longitudinal studies fulfilling the eligibility criteria. The sample sizes of the included studies ranged from n = 24 (Paananen, 2009; Speer & Meeks, 1985) to n = 285 (Lamont & Cross, 1994). The studies' contributions to the pooled effect size estimation differed, with weights ranging from approximately 2% (Paananen, 2009; Speer & Meeks, 1985) to 20% (Lamont & Cross, 1994). Effect sizes varied widely, $0 \leq d \leq 3.9$ with $0.14 \leq \upsilon \leq 0.41$, indicating substantial variance.

PsychOpen GOLD

Figure 1

Study Selection Flowchart

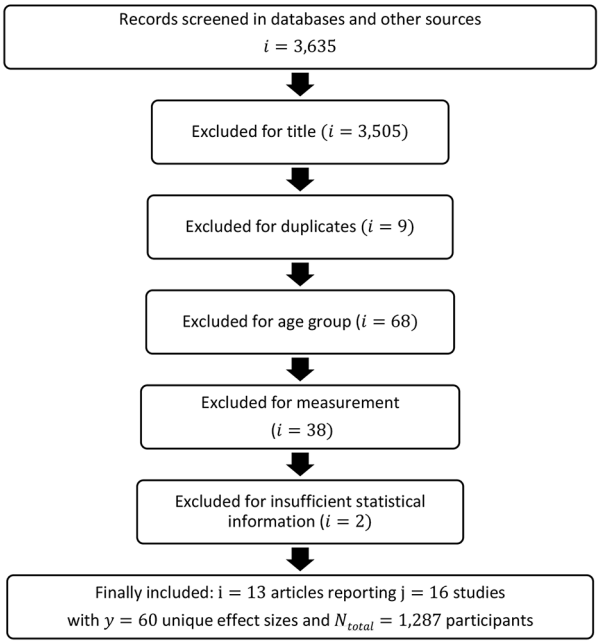


Table 1

Descriptive Statistics of Effect Sizes in Primary Studies and the Multi-Level Data Structure

	ID					
Author (Year)	**Article**	**Study**	**ES**	d	s^2	N
Krumhansl and Keil (1982)	1	1	1	2.35	0.24	28
Krumhansl and Keil (1982)	1	1	2	1.78	0.20	28
Krumhansl and Keil (1982)	1	1	3	0.00	0.14	28
Krumhansl and Keil (1982)	1	1	4	0.00	0.14	28
Krumhansl and Keil (1982)	1	1	5	0.00	0.14	28
Krumhansl and Keil (1982)	1	1	6	0.87	0.16	28
Krumhansl and Keil (1982)	1	1	7	1.10	0.16	28
Krumhansl and Keil (1982)	1	1	8	0.56	0.15	28
Krumhansl and Keil (1982)	1	1	9	0.00	0.14	28
Krumhansl and Keil (1982)	1	1	10	0.00	0.14	28
Krumhansl and Keil (1982)	1	1	11	3.90	0.41	28
Krumhansl and Keil (1982)	1	1	12	3.02	0.31	28

PsychOpen GOLD

	ID					
Author (Year)	**Article**	**Study**	**ES**	d	s^2	N
Krumhansl and Keil (1982)	1	1	13	0.72	0.15	28
Krumhansl and Keil (1982)	1	1	14	0.60	0.15	28
Krumhansl and Keil (1982)	1	1	15	1.24	0.17	28
Speer and Meeks (1985)	2	2	16	0.66	0.18	24
Cuddy and Badertscher (1987)	3	3	17	0.50	0.10	41
Cuddy and Badertscher (1987)	3	3	18	0.58	0.14	33
Schwarzer et al. (1993)	4	4	19	1.21	0.10	46
Schwarzer et al. (1993)	4	4	20	1.49	0.11	46
Schwarzer et al. (1993)	4	4	21	1.70	0.12	46
Lamont and Cross (1994)	5	5	22	0.69	0.02	285
Wilson and Wales (1995)	6	6	23	0.10	0.07	54
Paananen (2003)	7	7	24	0.61	0.17	24
Paananen (2003)	7	7	25	0.78	0.18	24
Paananen (2003)	7	7	26	0.53	0.17	24
Paananen (2003)	7	7	27	0.40	0.17	24
Paananen (2003)	7	7	28	0.49	0.17	24
Paananen (2003)	7	7	29	0.31	0.17	24
Paananen (2003)	7	7	30	0.30	0.17	24
Paananen (2003)	7	7	31	0.28	0.17	24
Paananen (2003)	7	7	32	0.43	0.17	24
Paananen (2003)	7	7	33	0.23	0.17	24
Schellenberg et al. (2005) I	8	8	34	0.68	0.19	23
Schellenberg et al. (2005) II	8	9	35	0.53	0.12	36
Schellenberg et al. (2005) III	8	10	36	0.43	0.09	44
Paananen (2009)	9	11	37	0.46	0.26	17
Paananen (2009)	9	11	38	-0.05	0.26	17
Paananen (2009)	9	11	39	-0.36	0.18	23
Paananen (2009)	9	11	40	0.62	0.18	23
Paananen (2009)	9	11	41	1.56	0.27	20
Paananen (2009)	9	11	42	0.78	0.22	20
Maier-Karius and Schwarzer (2011)	10	12	43	0.71	0.06	72
Maier-Karius and Schwarzer (2011)	10	12	44	1.23	0.15	42
James et al. (2012)	11	13	45	0.77	0.04	112
James et al. (2012)	11	13	46	0.11	0.04	112
James et al. (2012)	11	13	47	0.45	0.04	112
Matsunaga et al. (2020) I	12	14	48	0.69	0.09	48
Matsunaga et al. (2020) I	12	14	49	0.46	0.08	50
Matsunaga et al. (2020) I	12	14	50	-0.04	0.08	52
Matsunaga et al. (2020) I	12	14	51	0.29	0.09	48

PsychOpen GOLD

	ID					
Author (Year)	**Article**	**Study**	**ES**	d	s^2	N
Matsunaga et al. (2020) I	12	14	52	0.63	0.09	48
Matsunaga et al. (2020) II	12	15	53	-0.37	0.08	51
Matsunaga et al. (2020) II	12	15	54	0.18	0.09	44
Matsunaga et al. (2020) II	12	15	55	0.71	0.10	42
Matsunaga et al. (2020) II	12	15	56	-0.09	0.08	49
Matsunaga et al. (2020) II	12	15	57	0.67	0.08	52
Corrigall et al. (2022)	13	16	58	0.33	0.04	97
Corrigall et al. (2022)	13	16	59	-0.25	0.04	97
Corrigall et al. (2022)	13	16	60	0.82	0.04	97

The studies included participants from various age groups, ranging from age 6 (e.g., Krumhansl & Keil, 1982) to age 15 (Paananen, 2009). Adults were used as the treatment group in five studies (e.g., Matsunaga et al., 2020) because they were assumed to have fully developed IRTH, making them a suitable reference population. However, as suggested by our subsequent model comparison analyses, this assumption appears to be only partially accurate, as adults do not consistently demonstrate superior performance in tasks with higher cognitive demands. For a visualization of the correlation between the age of the treatment group, the age of the control group, and the effect sizes, see the scatter plot in Supplemental Material S9. Most studies used the probe tone technique for the dependent variable, either with a rating scale ($j = 5$) or by asking the participants to produce the most appropriate probe tone ($j = 1$). This includes studies using goodness-of-fit ratings for syntax violations and three studies using creative tasks, such as composition, improvisation, or harmonization (Table 2).

PsychOpen GOLD

Jahrbuch Musikpsychologie
2025, Vol. 33, Article e205
https://doi.org/10.5964/jbdgm.205

Table 2

Characteristics of Primary Studies

			n		Mean Age		Children with Formal Musical Training (%)	
ID	**Author (Year)**	**DV**	**Treat**	**Cont**	**Treat**	**Cont**	**Treat**	**Cont**
1	Krumhansl and Keil (1982), comp. 1	PT	14	14	8.5	6.5	71	43
1	Krumhansl and Keil (1982), comp. 2	PT	14	14	10.5	8.5	79	71
1	Krumhansl and Keil (1982), comp. 3	PT	14	14	20	10.5	86	79
2	Speer and Meeks (1985)[a]	PT	12	12	10	7	42	33
3	Cuddy and Badertscher (1987), comp. 1	PT	21	20	8.5	6.5	43	45
3	Cuddy and Badertscher (1987), comp. 2	PT	12	21	10.5	8.5	42	43
4	Schwarzer et al. (1993)[a]	PTP	20	26	20	9	25	23
5	Lamont and Cross (1994)[b]	PT	285	—	6-9	—	n/a	n/a
6	Wilson and Wales (1995)	Comp	36	37	9	7	42	62
7	Paananen (2003), comp. 1	Imp	12	12	8.5	6.5	n/a	n/a
7	Paananen (2003), comp. 2	Imp	12	12	10.5	8.5	n/a	n/a
8	Schellenberg et al. (2005) I	RT	13	10	10.5	6.5	100	0
9	Schellenberg et al. (2005) II	RT	19	17	10.5	7.5	47	41
10	Schellenberg et al. (2005) III	RT	22	22	10.5	7.5	59	50
11	Paananen (2009), comp. 1	Harm	12	11	8.5	6.5	n/a	n/a
11	Paananen (2009), comp. 2	Harm	12	11	10.5	8.5	n/a	n/a
11	Paananen (2009), comp. 3	Harm	8	12	14.5	10.5	n/a	n/a
12	Maier-Karius and Schwarzer (2011), comp. 1	PT	32	40	9.5	6.5	25	18
15	Maier-Karius and Schwarzer (2011)[a], comp. 2	PT	10	32	20	9.5	100	25
13	James et al. (2012)[b]	GoF	112	—	6-10	—	n/a	n/a
14	Matsunaga et al. (2020) I, comp. 1	GoF	24	24	9	7	0	08

PsychOpen GOLD

ID	Author (Year)	DV	*n*		Mean Age		Children with Formal Musical Training (%)	
			Treat	Cont	Treat	Cont	Treat	Cont
14	Matsunaga et al. (2020) I, comp. 2	GoF	26	24	10.5	9	0	0
14	Matsunaga et al. (2020) I, comp. 3	GoF	28	26	12.5	10.5	0	0
14	Matsunaga et al. (2020) I, comp. 4	GoF	20	28	14	13.5	0	0
14	Matsunaga et al. (2020) I, comp. 5	GoF	28	28	20	14	0	0
15	Matsunaga et al. (2020) II, comp. 1	GoF	25	26	9	7	0	0
15	Matsunaga et al. (2020) II, comp. 2	GoF	19	25	11	9	0	0
15	Matsunaga et al. (2020), II, comp. 3	GoF	23	19	13	11	0	0
15	Matsunaga et al. (2020) II, comp. 4	GoF	26	23	15	13	0	0
15	Matsunaga et al. (2020) II, comp. 5	GoF	26	26	20	15	0	0
16	Corrigall et al. (2022)	RT, GoF	49	48	6.5	10.5	55	45

Note. comp. = age group comparison in case of multi-arm studies; 1 = youngest group; 2 = next oldest group, etc., up to comp. 5; DV = dependent variable; PT = probe tone rating; PTP = probe tone production (participants played the probe tone on a keyboard); Imp = improvisation; Comp = composition; Harm = harmonization; RT = response time; GoF = goodness-of-fit ratings; Treat = treatment group (older participants); Cont = control group (younger children); *n* = number of participants in each group; n/a in the columns of musical training: the available information is insufficient to reproduce the number of musically trained participants in each subgroup. Formal musical training is operationalized differently in each study.

[a]indicates sudies, in which the mean age for adults was set to 20 years due to insufficient information, in detail, Krumhansl and Keil (1982, p. 246) reported only undergraduate or graduate students as adults, Maier-Karius and Schwarzer (2011, p. 172) graduate students, and Schwarzer et al. (1993, p. 77) provided an age range of 19–45 years. [b]provided only a global effect size for primary school students; therefore, the age of the youngest and oldest participants were coded.

Jahrbuch Musikpsychologie
2025, Vol. 33, Article e205
https://doi.org/10.5964/jbdgm.205

PsychOpen GOLD

The control variables differed significantly across studies. Some researchers, such as Krumhansl and Keil (1982), did not explicitly control for prior formal musical training. In contrast, others, such as Corrigall et al. (2022), treated it as a key variable. The participants with formal musical training in each sample ranged from none (coded as 0; Matsunaga et al., 2020) to all (coded as 1; Maier-Karius & Schwarzer, 2011). Three studies reported the percentage of musically trained participants only across the sample: Paananen (2003), 13.89%; Paananen (2009), 13.63%; and James et al. (2012), 16.96%. Formal musical training could not be included as a moderator in the statistical model, owing to limited data and a lack of an apparent long-term sampling practice considering these variables (Harrer et al., 2021)

Study quality was evaluated using various criteria (see Supplemental Material S2), with an average score of M = 19.3 (SD = 2.3) out of 43. Overall, study quality was relatively homogeneous; however, only one study reported psychometric quality criteria (Wilson & Wales, 1995).

Bayesian Three-Level Meta-Analysis

We initially estimated a four-level meta-analysis to aggregate effect sizes, as two articles (Matsunaga et al., 2020; Schellenberg et al., 2005) reported independent studies, suggesting potential within-article heterogeneity. However, although a comparison between the initial four-level model and a three-level meta-analysis model showed that the Bayes factor, BF_{10} = 1.92, provided insufficient evidence to decisively favor either model (for a general discussion, see van Doorn et al., 2021), the three-level model was ultimately chosen because of its slightly better data fit and based on the principle of parsimony (Vandekerckhove et al., 2015). The model comparison did not support further simplifying the model to a random-effects structure[4].

Before interpreting the results, we confirmed model convergence for our three-level structure by conducting posterior predictive checks and examining the $\widehat{R}$-values of the parameter estimates (Harrer et al., 2021). All $\widehat{R}$-values $\left(\widehat{R} \leq 1.01\right)$ indicated successful convergence (Bürkner, 2017). Additionally, visual inspection confirmed that the posterior distributions aligned with the initial unimodal normal distribution, supporting our assumption of normality (see Appendix Figure A2-1). Sensitivity analyses using different weakly informed priors further support the robustness assumption of our results (Table 3).

4) A Bayes factor comparing both models provided strong evidence in favor of the three-level model (BF_{10} = 33,283.83).

PsychOpen GOLD

Jahrbuch Musikpsychologie
2025, Vol. 33, Article e205
https://doi.org/10.5964/jbdgm.205

Table 3

Prior and Posterior Model Parameters of Three-Level Meta-Analysis Model and Sensitivity Analyses

	Priors of the Three-Level Model		Alternatives in Prior Choice		
	Prior 1	**Prior 2**	**Prior 3**	**Prior 4**	
Model Parameter	**N(0, 1)**	**HalfCauchy(0, 0.5)**	**N(0, 0.5)**	***t*(3, 0, 2.5)**	**Conclusions**
Intercept (μ)	*Mdn* = 0.57 (0.37, 0.77)	—	*Mdn* = 0.56 (0.36, 0.75)	*Mdn* = 0.58 (0.37, 0.78)	Small, but practically negligible difference for both estimates. Prior 3 goes along with a slightly different, but insignificant shrinkage of the point estimates towards the center of the prior)
Between-study (τ_1)	—	*Mdn* = 0.20 (0.00, 0.40)	*Mdn* = 0.20 (0.04, 0.40)	*Mdn* = 0.21 (0.04, 0.42)	Minor, but practically negligible differences between both parameter estimates. Prior 4 produced a slightly wider 95% credible interval, reflecting marginally more uncertainty.
Within-study (τ_2)	—	*Mdn* = 0.43 (0.27, 0.60)	*Mdn* = 0.43 (0.28, 0.60)	*Mdn* = 0.45 (0.27, 0.61)	Minor, but practically negligible variation across priors.

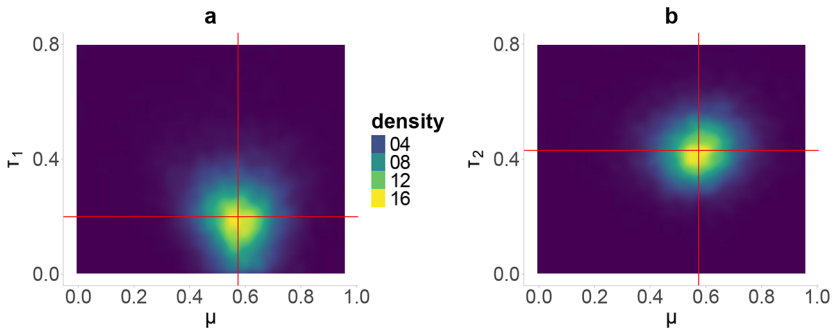
Note. Posterior medians (*Mdn*) are used as point estimates; 95% credible intervals are presented in parentheses; for Priors 3 and 4, half-distributions were used for modeling τ_1 and τ_2.

We assessed publication bias using a Bayesian Egger test, which indicated a significant intercept (β_1 = 7.46, 95% *CI* [3.12, 11.67]), suggesting potential publication bias. However, previous studies have recommended conducting this approach on a corpus of at least 30 studies because asymmetry tests for standardized mean differences are prone to inflated Type I error rates (Pustejovsky & Rodgers, 2019; Renkewitz & Keiner, 2019). A visual inspection of the funnel plot (see Appendix Figure A2-2), showing the standard error of the reported effect sizes as a function of their magnitude, suggests the potential for slight asymmetry; however, this is primarily driven by Krumhansl and Keil (1982; Study ID 1). Thus, we found no significant evidence of general publication bias caused by selective reporting.

We further explored potential sources of bias by examining the model-based relationships between μ and τ_1 (Figure 2a), and μ and τ_2 (Figure 2b). Although visual inspection suggested only marginal associations, Pearson correlations revealed statistically significant but negligible effects in both cases: μ and τ_1, $r(11998) = -.09$, $p < .001$, 95% *CI* (-.11, -.07); μ and τ_2, $r(11998) = .08$, $p < .001$, 95% *CI* (.06, .10). These results indicated that any potential relationship between effect sizes and heterogeneity is minimal, thus providing little evidence of systematic publication bias.

Figure 2

Heatmap of the Joint Posterior Distribution of Model Parameters μ (x-axis) and Standard Deviations τ (y-axis)



Note. Figure 2a depicts the relationship between μ and τ_1 (between-study standard deviation). Figure 2b shows the relationship between μ and τ_2 (within-study standard deviation). The color gradient reflects the density scale, with higher values (lighter regions) indicating higher density and lower values (darker regions) indicating lower density. Red lines in each plot represent the maximum likelihood point estimates of both parameters.

Based on the observed data, hierarchical model, and prior choices, we produced a point estimation for the pooled medium-sized effect (Mdn_d = 0.57). As indicated by the credible intervals of the pooled effect size (Table 3), we could further conclude that the general so-called "true" effect size—expressed as Cohen's *d*—reflects a medium to nearly large difference in IRTH sensitivity between younger and older participants. Moreover, the

PsychOpen GOLD

95% credible interval for the effect lies entirely above zero (0.37, 0.77)[5] along with decisive evidence against a point-null hypothesis (BF_{10} = 11999). We further assess whether the age-related increase in IRTH sensitivity exceeds a negligible magnitude, that is, whether it is neither trivially different from zero nor within the range of a small effect $\left(H_0{:}0 \leq |d| < 0.2,\ BF_{01} > 10\right)$. Therefore, we used Bayes factors to quantify the evidence that the result exceeded the small-effect threshold $\left(H_1{:}\ |d| \geq 0.2,\ BF_{10} > 10\right)$ by using a Bayesian model comparison approach. Our analyses revealed a Bayes Factor providing decisive support for the alternative assumption (BF_{10} = 1332.33) indicating that the age-related increase in IRTH sensitivity is unlikely to be marginal. Instead, it should be best interpreted as a substantively meaningful skill-development effect with a high probability of at least medium-to-strong magnitude. We further illustrated the cumulative posterior evidence by plotting the empirical cumulative distribution functions for μ, τ_1 and τ_2 (see Appendix Figure A2-3) to facilitate the interpretation of the probability that these parameters exceed meaningful thresholds.

Furthermore, based on the hierarchical model and its estimated parameters, we revealed a model with parameter estimates as solution in which at least one effect size estimate from each study falls within the 95% credible interval of the pooled effect size (Figure 3). At first glance, Schwarzer et al. (1993) appear to be an exception; however, although the point estimates of effect sizes in their study are predicted to fall well outside the 95% credible interval of the pooled effect size, their respective 95% credible intervals still intersect with this range.

Although the vast majority of the effect sizes reported in investigated studies fall within the expected 50–80% credible intervals according to our hierarchical model (Figure 4), some observed effect sizes reported by Krumhansl and Keil (1982) deviate from this pattern. Specifically, the results of two studies (ID 11 and ID 12) exhibit exceptionally high effect sizes. Although our model considers these values possible, they are highly unlikely from a statistical standpoint. For example, according to our model, the probability that the estimate for the effect size for ID 12 in Krumhansl and Keil (1982), denoted as $\theta_{1,12}$ in the posterior samples, is greater than or equal to the observed effect size $\left(\mathrm{ES}_{1,12} = 3.02\right)$ is relatively low at 2.1%, $P\left(\theta_{1,12} \geq \mathrm{ES}_{1,12} \middle| y, M\right) = 0.021$. In contrast, the observed effect size for ID 11 has an even lower probability of occurring, at less than 1% $\left(P\left(\theta_{1,11} \geq \mathrm{ES}_{1,11} \middle| y, M\right) = 0.005\right)$.

5) The 95% credible interval (95% CrI) is a highest density interval, i.e., the smallest possible interval containing the probability mass of 95% of the most probable values in the posterior parameter distribution. It represents the region where every point has a higher probability density than any point outside, reflecting the uncertainty in the estimate (van Doorn et al., 2021, p. 822).

PsychOpen GOLD

Figure 3

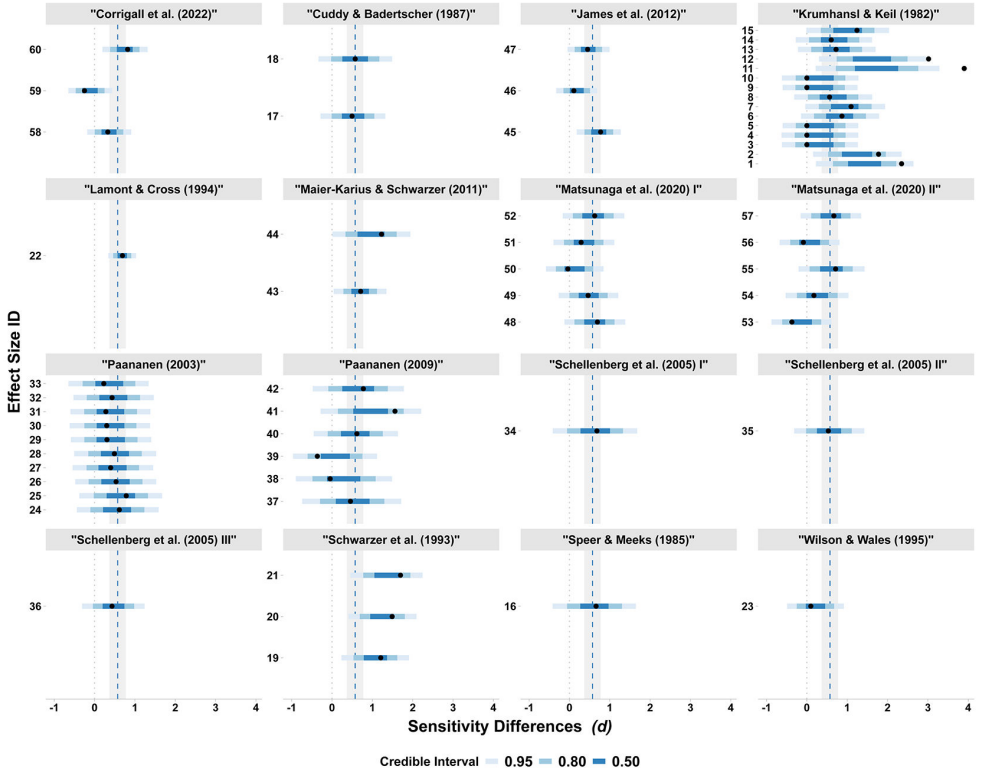
Forest Plot of the Bayesian Three-Level Model With Estimated Effect Sizes of Individual Studies and the Pooled Effect Size

Note. The densities represent their respective posterior distributions. Medians serve as point estimates for the effect sizes. Blue shading highlights estimated effect sizes and parts of the posterior distributions that fall within the 95% credible interval of the pooled effect size.

PsychOpen GOLD

Figure 4

Model-Based Probability Predictions for the Occurrence of Each Effect Size



Note. Whereas the vast majority of effect sizes have a high probability of occurrence based on the model-based expectations, some values have extremely low statistical probability, such as ID 11 and ID 12.

The multilevel model revealed significant heterogeneity, divided into between- and within-study heterogeneity. Although the median between-study variance, $\tau_1^2 = 0.04,\ 95\%\ CrI\ (0.00,\ 0.16)$, accounted for only 18% of the total variance, the majority of the heterogeneity (82%) could be attributed to within-study variance, $\tau_2^2 = 0.18,\ 95\%\ CrI\ (0.07,\ 0.36)$.

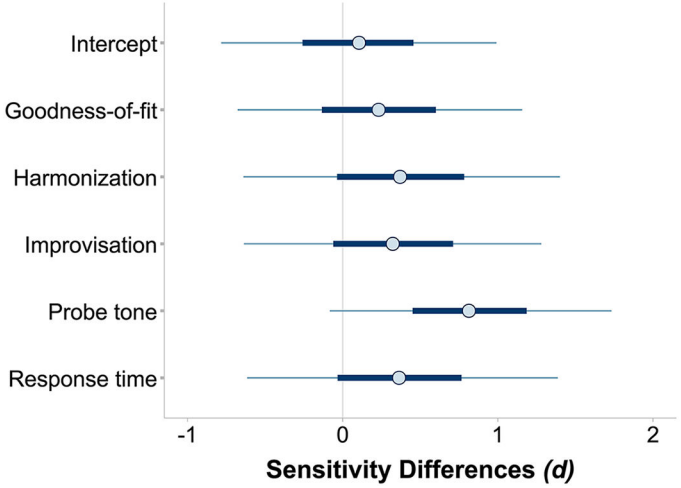
However, differences in age-group compositions across studies prevented us from answering RQ2 using this analytical approach, which is discussed below. Furthermore, RQ3 could only be addressed in an exploratory manner, as the limited number of studies and inconsistent reporting of relevant sample characteristics hindered formal statistical testing. Table 2 summarizes the proportions of musically trained and untrained individuals across the included samples.

PsychOpen GOLD

We addressed RQ4 by examining whether differences in IRTH measurement type could explain the variability in effect sizes. The central goal of a meta-analysis is to account for heterogeneity in the data (Borenstein, 2019). Therefore, we followed Corrigall et al. (2022) and introduced a categorical moderator representing six measurement types: probe tone, response time, goodness-of-fit rating, composition, improvisation, and harmonization. The analysis revealed largely overlapping 95% credible intervals and only small differences in posterior means, suggesting that the measurement type had no systematic effect (Figure 5). However, comparing model fit using R^2 values (Hayes, 2022) showed a slight improvement in the moderated model, R^2 = .51, 95% *CrI* (.30, .72), compared to the model without a moderator, R^2 = .48, 95% *CrI* (.27, .70). Although the Bayes factor of BF_{10} = 38.80 indicated a strong statistical preference for the moderated model (van Doorn et al., 2021), the practical significance of this improvement remained uncertain, given the modest differences in parameter estimates. Therefore, we directly addressed RQ4 by comparing the response time measures with all other operationalizations. These comparisons yielded weak-to-moderate evidence against a systematic difference for goodness-of-fit ratings (BF_{10} = 0.48), harmonization (BF_{10} = 1.00), and improvisation (BF_{10} = 0.84) in lower estimates for response time measures compared to probe tone ratings (BF_{10} = 12.99). Overall, the results indicated smaller effect sizes for response time measurements in probe tone ratings only, with no consistent reduction in effect sizes for response time measurements.

Figure 5

Posterior Distributions of the Intercept and Moderator Levels of the Bayesian Three-Level Moderated Meta-Analysis



Note. Effect sizes are standardized mean differences (*d*). Thin error bars are 50% *CrI* of *d*. Thick error bars indicate 95% *CrI* of *d*.

PsychOpen GOLD

Taken together, these results suggest that response time measurements do not consistently yield lower IRTH sensitivity estimates compared to more explicit operationalizations. Therefore, the current evidence does not support RQ4, except for a notable difference in probe tone ratings.

The observed within- and between-study heterogeneity was likely caused by specific artifacts, such as sample and characteristics. However, we cannot pinpoint these factors more precisely because of the limited information in the primary studies. However, certain possibilities were excluded based on the model comparison analysis presented below.

Thus far, we have reported evidence of a medium-to-strong age-related increase in IRTH sensitivity, although the high variance somewhat obscures this trend. However, these analyses do not clarify whether age-specific learning gains are present, particularly (1) whether acquisition continues beyond age 9 and (2) whether there are sensitive periods during school years that would be reflected in a non-linear learning trajectory. Thus, we conducted further analyses to address these uncertainties.

Model Comparison Analysis

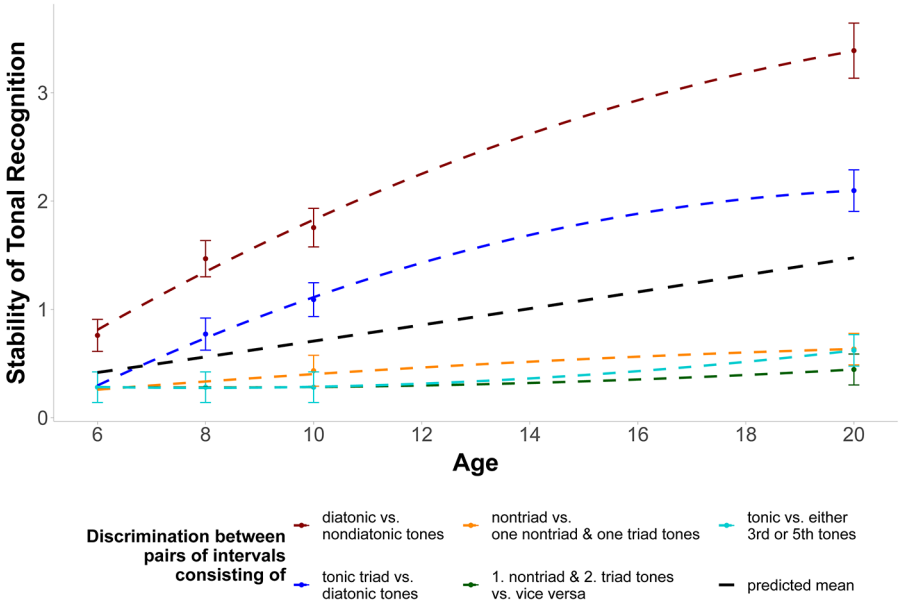
Based on the *AIC*, *BIC*, and *RMSEA* model comparison criteria, the non-linear mixed model best fits Krumhansl and Keil's (1982) data (see Appendix Table A1-1), suggesting no linear or sigmoidal increase in IRTH sensitivity between ages 6 and 20. Figure 6 shows the age-related development of various IRTH sensitivity outcomes according to the non-linear mixed model. Based on the maximum increase in IRTH sensitivity at age 20 ($k = 0.08$), we concluded that IRTH continues to develop into adulthood and is not fully developed by age 9.

Furthermore, the distinct trajectories shown in Figure 6 suggest that outcomes may strongly depend on specific task characteristics, even within a single-measurement paradigm. Although the differentiation between pairs of diatonic and non-diatonic tones, as well as the tonic triad versus other diatonic tones, shows a steep increase, more subtle IRTHs, such as the preference for the tonic over other tonic triad tones, show little to no increase, even in adults.

PsychOpen GOLD

Figure 6

Development of IRTH Modelled as a Non-Linear Function of Age



Note. Standardized mean differences (*d*) as indicator of tonal recognition stability based on reanalyzed cross-sectional data of Krumhansl and Keil (1982, pp. 247–248) as measured with the two sample tones probe tone paradigm (see section procedure for explanation).

Discussion

Our study aimed to synthesize 40 years of research examining the impact of age-related experience with Western tonal music on IRTH acquisition in school-aged children. Additionally, we explored sources of variance both within and between studies to better understand the heterogeneity in study outcomes.

In a first approach, we aggregated eligible studies in a Bayesian three-level meta-analysis to precisely quantify the mean increase in IRTH sensitivity between younger and older participants. On average, the results indicated a moderate increase in IRTH acquisition during one's school years, depending on the cognitive demands of the tasks measuring IRTH sensitivity. This finding has significant educational implications. The observed medium effect size supports the implementation of targeted programs (e.g., Government of the UK, Department of Education, 2021) to enhance tonal sensitivity in school-aged children, as this is sufficiently substantial to be practically relevant for educational initiatives (Hattie, 2012). Future research should further investigate these

PsychOpen GOLD

acquisitional trajectories and explore potential critical periods beyond the age range covered in our study (Hargreaves & Lamont, 2017).

Subsequently, we conducted a model comparison of single-study results to explore the timing and potential shapes of learning trajectories within IRTH acquisition as a function of task characteristics. The findings provide converging evidence against early closure models that assume a fixed endpoint of acquisition by age 7 or 9 (Gembris, 2017b; Gordon, 2012). Our findings indicate that IRTH sensitivity instead continues to develop through adolescence and into adulthood, supporting gradual or open-ended models of acquisition (e.g., Hargreaves, 1996) and aligning with neurocognitive and educational frameworks of lifelong learning (Altenmüller, 2022; Mack et al., 2025). We also observed improvements among participants without formal musical training, underscoring the role of enculturation and mere exposure (Demorest & Morrison, 2016). Furthermore, data from participants with ongoing formal training—such as the adult volunteers in Krumhansl and Keil (1982), who had 7.8 years of musical training on average—indicate that IRTH performance remains trainable beyond early childhood. These patterns suggest a skill development process shaped not only by exposure but also by the cognitive demands of the task. Higher task complexity may require a level of expertise that emerges only through extended, potentially formalized practice. Consequently, the observed plateau effects likely reflect task-specific thresholds rather than universal developmental limits.

A more detailed analysis using a model comparison approach revealed that a task-specific non-linear model best captured the developmental trajectory, a pattern also observed in other musical contexts, such as the development of melody perception (Lin, 2023). Notably, our analysis revealed substantial differences in IRTH developmental trajectories, which varied not only as a non-linear function of age but also as a function of specific task characteristics. Matsunaga et al. (2020) also reported variations across tasks, who found that Japanese children recognized Western diatonic tones at age 7 but identified the tonic only at age 13. In contrast, implicit measures of IRTH demonstrated sensitivity to the tonic in ages 6 to 7 (Corrigall et al., 2022; Schellenberg et al., 2005), suggesting that implicit processes may precede explicit ones in tonal development (Corrigall et al., 2022). Neuroscientific (Corrigall & Trainor, 2014) and behavioral studies using implicit measures (Trainor & Trehub, 1992) further support the notion that some tonal abilities might develop earlier than others. Our meta-analysis, which focused primarily on explicit tasks, aligns with research showing later acquisition in explicit tonal processing.

However, data from all reviewed studies currently lack sufficient evidence to determine the correlative relationship between the various operationalizations used or whether they can be attributed to a single latent variable (IRTH) to which all prior operationalizations conceptually refer. To date, no statistical examination has been conducted to determine whether all operationalizations capture the same latent variable, assuming

PsychOpen GOLD

that IRTH follows a general factor model. Alternatively, these operationalizations may represent dimensions of a composite factor model or discrete partially uncorrelated constructs associated with different latent variables. A test-theoretical approach would be valuable to address this question across age groups (e.g., Bond & Fox, 2015), enabling the subsequent examination of developmental trajectories of IRTH in a longitudinal study. Owing to insufficient statistical information in the primary studies, RQ3 regarding the effects of formal musical training on IRTH acquisition could not be answered fully based on the available evidence. However, it can be inferred from our findings that although IRTH sensitivity develops in children without formal musical training (e.g., Matsunaga et al., 2020), formal musical training may enhance explicit judgments (e.g., $d = 1.23$ for musical experts in Maier-Karius & Schwarzer, 2011). However, it has minimal impact on implicit response times (e.g., $d = 0.68$ falling within the 95% CrI of the modeled mean in Schellenberg et al., 2005). This pattern is consistent with Corrigall et al.'s (2022) findings and is further supported by studies employing productive IRTH operationalizations (e.g., Guilbault, 2009; Wilson & Wales, 1995). However, several studies have reported no advantage of formal musical training for goodness-of-fit ratings (James et al., 2015; Schellenberg et al., 2005; Stalinski & Schellenberg, 2010) or observed improvements in implicit measures such as brain function in musically trained participants (Koelsch et al., 2005; Magne et al., 2006; Putkinen et al., 2014; Wehrum et al., 2011).

Although the limited data prevented us from disentangling the individual contributions of age, informal and formal musical training, and their interactions, our analyses underscore the need for further research in this area. The recurrent absence of data emphasizes the importance of transparent and sustainable research data management (Eerola, 2025), including practices such as verbatim reporting of instructions, providing original stimuli, reporting means and standard deviations for each condition and treatment group (or supplying raw data), and specifying test quality criteria (e.g., test-retest correlation for repeated measurements). By addressing these aspects, future studies will enable meta-analyses such as ours to be conducted more broadly and provide more precise estimates.

The following limitations should be considered when interpreting the results. Generally, meta-analyses inherit the methodological limitations of primary studies. For example, all included studies used cross-sectional quasi-randomized or observational study designs. Although common in this field (Boutron et al., 2022), these methods may introduce additional variance (Murad et al., 2016). Another example is the predominantly low statistical power of the primary studies (Ellis, 2010). Although meta-analytic weighting procedures account for sampling errors by assigning lower weights to less precise estimates, studies with small samples remain more susceptible to extreme or unstable effect size estimates owing to higher random variability (Harrer et al., 2021). Therefore, a high proportion of such studies may increase heterogeneity, particularly if their estimates are systematically biased or selectively reported. This highlights the

PsychOpen GOLD

importance of adequately designed and reported primary studies and calls for caution when interpreting highly variable or inconsistent findings stemming from small samples. From a conceptual standpoint, a limitation of our study lies in the treatment of age as an independent variable based on its frequent reporting in several studies. Age influences various cognitive factors, such as maturation, enculturation, exposure, and musical skill acquisition, that are likely to affect IRTH acquisition (Demorest & Morrison, 2016; Halford, 2014; Hannon & Trainor, 2007; Zajonc, 2001). However, the specific contributions of these factors remain unclear in both the primary studies and our own analyses. For example, formal musical training is known to enhance IRTH (Corrigall & Trainor, 2009; Kraus & Chandrasekaran, 2010; Müllensiefen, 2022; Virtala et al., 2012) but likely interacts with age and informal musical activities (Lamont, 1998). Future research should aim to disentangle the contributions of these factors and clarify their interactions to better understand IRTH acquisition. Another limitation is the substantial heterogeneity observed in the Bayesian three-level meta-analysis, which may have obscured the overall effect. This heterogeneity can be primarily attributed to Krumhansl and Keil (1982).

In summary, although our research question regarding smaller effect sizes for response time measurements compared to other more explicit measurements remains partially unanswered, with probe tone ratings being the exception, the model comparison analysis of Krumhansl and Keil (1982) indicates task-specific differences. Our findings underscore the pivotal influence of task characteristics on IRTH measurements reported over the past 40 years. This aligns with prior research highlighting performance variations stemming from differences in probe tone instructions (Kristop et al., 2020) and supports insights gained from re-analyses of measurement instruments (Platz et al., 2022). Reviewing the knowledge gained from 40 years of research in a structured manner is worthwhile because it provides specific insights for the profitable continuation of the research tradition and general findings for appropriately handling data that enable fruitful re-analyses. This discussion highlights key challenges in IRTH research and underscores the importance of employing validated methodological approaches to accurately model and interpret data. Future studies should address these limitations and focus on developing more precise and comprehensive models based on valid measures to better capture the complex dynamics of IRTH development.

PsychOpen GOLD

Funding: The authors have no funding to report.

Acknowledgments: We would like to thank Editage (www.editage.com) for English language editing.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Hanna Mütze*—Conceptualization | Methodology | Investigation | Data curation | Formal analysis | Visualization | Writing – original draft. *Friedrich Platz*—Conceptualization | Supervision | Validation | Writing – review & editing. *Veronika Busch*—Conceptualization | Supervision | Validation | Writing – review & editing | Project administration.

Ethics Statement: No ethical issues and/or ethics approvals need to be disclosed.

Data Availability: For this article, data and codebook are freely available (see Mütze et al., 2025a).

Supplementary Materials

For this article, the following Supplementary Materials are available:

- the raw data of the effect sizes, additionally coded variables of the meta-analysis ("Supplemental Material S4"), the raw data of the model comparison analysis ("Supplemental Material S7") as well as the associated codebooks (see Mütze et al., 2025a)
- the R markdown with the code of the associated statistical analyses of the meta-analysis ("Supplemental Material S5") as well as the R markdown with the code for the associated statistical analysis of the model comparison analysis ("Supplemental Material S8") (see Mütze et al., 2025b)
- the review protocol (referred to as "Supplemental Material S1" in the manuscript), the coding protocol ("Supplemental Material S2"), the methods of effect size calculations ("Supplemental Material S3"), and the formal description of the mathematical model ("Supplemental Material S6") (see Mütze et al., 2025c)
- the Scatterplot ("Supplemental Material S9") (see Mütze et al., 2025d)

Index of Supplementary Materials

Mütze, H., Platz, F., & Busch, V. (2025a). *Supplementary materials to "The acquisition of tonal hierarchies in western music during school years: A re-analysis of 40 years of research"* [Data, codebook]. PsychArchives. https://doi.org/10.23668/psycharchives.21181

Mütze, H., Platz, F., & Busch, V. (2025b). *Supplementary materials to "The acquisition of tonal hierarchies in western music during school years: A re-analysis of 40 years of research"* [Code]. PsychArchives. https://doi.org/10.23668/psycharchives.21336

Mütze, H., Platz, F., & Busch, V. (2025c). *Supplementary materials to "The acquisition of tonal hierarchies in western music during school years: A re-analysis of 40 years of research"* [Protocols, additional materials]. PsychArchives. https://doi.org/10.23668/psycharchives.21183

PsychOpen GOLD

Mütze, H., Platz, F., & Busch, V. (2025d). *Supplementary materials to "The acquisition of tonal hierarchies in western music during school years: A re-analysis of 40 years of research"* [Scatterplot]. PsychArchives. https://doi.org/10.23668/psycharchives.21184

References

Altenmüller, E. (2022). Brain mechanisms of music learning and performing. In G. E. McPherson (Ed.), *The Oxford handbook of music performance: Development and learning, proficiencies, performance practices, and psychology* (Vol. 2, pp. 157–178). Oxford University Press.https://doi.org/10.1093/oxfordhb/9780190056285.013.24

Asztalos, K., & Csapo, B. (2017). Development of musical abilities: Cross-sectional computer-based assessments in educational contexts. *Psychology of Music, 45*(5), 682–698. https://doi.org/10.1177/0305735616678055

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Taylor & Francis Group.

Borenstein, M. (2019). Heterogeneity in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 453–470). Russell Sage Foundation.

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–243). Russell Sage Foundation.

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., & Hróbjartsson, A. (2022). Chapter 7: Considering bias and conflicts of interest among the included studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Chumpston, T. Li, M. J. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.3, updated February 2022). Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/archive/v6.3

Brehmer, F. (1925). *Melodieauffassung und melodische Begabung des Kindes* [Melodic perception and melodic giftedness in the child]. Barth.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80,* 1–28. https://doi.org/10.18637/jss.v080.i01

Corrigall, K. A., & Schellenberg, E. G. (2016). Music cognition in childhood. In S. Hallam, I. Cross, & M. H. Thaut (Eds.), *The child as musician: A handbook of musical development* (2nd ed., pp. 81–101). Oxford University Press.

Corrigall, K. A., Tillmann, B., & Schellenberg, E. G. (2022). Measuring children's harmonic knowledge with implicit and explicit tests. *Music Perception, 39*(4), 361–370. https://doi.org/10.1525/mp.2022.39.4.361

PsychOpen GOLD

Corrigall, K. A., & Trainor, L. J. (2009). Effects of musical training on key and harmonic perception. *Annals of the New York Academy of Sciences, 1169*(1), 164–168. https://doi.org/10.1111/j.1749-6632.2009.04769.x

Corrigall, K. A., & Trainor, L. J. (2014). Enculturation to musical pitch structure in young children: Evidence from behavioral and electrophysiological methods. *Developmental Science, 17*(1), 142–158. https://doi.org/10.1111/desc.12100

Cuddy, L. L., & Badertscher, B. (1987). Recovery of the tonal hierarchy: Some comparisons across age and levels of musical experience. *Perception & Psychophysics, 41*(6), 609–620. https://doi.org/10.3758/BF03210493

Cui, A.-X. (2019). *Characterizing statistical learning of music* [Dissertation, Queen's University]. ProQuest Dissertations Publishing. https://www.proquest.com/openview/89fa1699a21bf11f4c790d98c5fcab6a/1?pq-origsite=gscholar&cbl=18750&diss=y

Demorest, S. M., & Morrison, S. J. (2016). Quantifying culture: The cultural distance hypothesis of melodic expectancy. In J. Y. Chiao, S.-C. Li, R. Seligman, & R. Turner (Eds.), *The Oxford handbook of cultural neuroscience* (pp. 183–194). Oxford University Press.

Eerola, T. (2025). Prevalence of transparency and reproducibility-related research practices in music psychology (2017–2022). *Musicae Scientiae, 29*(2), 385–399. https://doi.org/10.1177/10298649241300885

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

Gardner, H. (1973). *The arts and human development: A psychological study of the artistic process.* John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2015). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Gembris, H. (2017a). Entwicklungstheorien und empirische Befunde [Developmental theories and empirical results]. In H. Gembris (Ed.), *Grundlagen musikalischer Entwicklung und Begabung* (5th ed., pp. 233–265). Wißner.

Gembris, H. (2017b). *Grundlagen musikalischer Begabung und Entwicklung [Foundations of musical giftedness and development]* (5th ed.). Wißner.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gordon, E. E. (2012). *Learning sequences in music: A contemporary music learning theory* (8th ed.). Gia Publications.

Government of the UK, Department of Education. (2021). *National curriculum in England: Music programmes of study.* https://www.gov.uk/government/publications/national-curriculum-in-england-music-programmes-of-study/national-curriculum-in-england-music-programmes-of-study

PsychOpen GOLD

Guilbault, D. M. (2009). The effects of harmonic accompaniment on the tonal improvisations of students in first through sixth grade. *Journal of Research in Music Education, 57*(2), 81–91. https://doi.org/10.1177/0022429409337201

Halford, G. S. (2014). *Children's understanding: The development of mental models*. Psychology Press.

Hanna-Pladdy, B., & MacKay, A. (2011). The relation between instrumental musical activity and cognitive aging. *Neuropsychology, 25*(3), 378–386. https://doi.org/10.1037/a0021895

Hannon, E. E., & Trainor, L. J. (2007). Music acquisition: Effects of enculturation and formal training on development. *Trends in Cognitive Sciences, 11*(11), 466–472. https://doi.org/10.1016/j.tics.2007.08.008

Hargreaves, D. J. (1996). The development of artistic and musical competence. In I. Deliege & J. A. Sloboda (Eds.), *Musical beginnings: Origins and development of musical competence* (pp. 145–170). Oxford University Press.

Hargreaves, D. J., & Lamont, A. (2017). Cognition, perception, and learning. In D. J. Hargreaves & A. Lamont (Eds.), *The psychology of musical development* (1st ed., pp. 14–57). Cambridge University Press. https://doi.org/10.1017/9781107281868

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis in R: A hands-on guide*. Chapman and Hall/CRC. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/bayesianma.html

Hattie, J. (2012). *Visible learning*. Routledge.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

Hedges, L. V. (2019). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 281–298). Russell Sage Foundation.

Hyer, B. (2021). *Tonality* [Online Dictionary]. Grove Music Online. https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000028102

James, C. E., Cereghetti, D. M., Tribes, E. R., & Oechslin, M. S. (2015). Electrophysiological evidence for a specific neural correlate of musical violation expectation in primary-school children. *NeuroImage, 104*, 386–397. https://doi.org/10.1016/j.neuroimage.2014.09.047

James, C. E., Dupuis-Lozeron, E., & Hauert, C.-A. (2012). Appraisal of musical syntax violations by primary school children: Effects of age and practice. *Swiss Journal of Psychology, 71*(3), 161–168. https://doi.org/10.1024/1421-0185/a000084

Janata, P., & Reisberg, D. (1988). Response-time measures as a means of exploring tonal hierarchies. *Music Perception, 6*(2), 161–172. https://doi.org/10.2307/40285423

Jentschke, S., Koelsch, S., & Friederici, A. D. (2005). Investigating the relationship of music and language in children: Influences of musical training and language impairment. *Annals of the New York Academy of Sciences, 1060*(1), 231–242. https://doi.org/10.1196/annals.1360.016

Jonaitis, E. M., & Saffran, J. R. (2009). Learning harmony: The role of serial statistics. *Cognitive Science, 33*(5), 951–968. https://doi.org/10.1111/j.1551-6709.2009.01036.x

PsychOpen GOLD

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., & Schlaug, G. (2005). Adults and children processing music: An fMRI study. *NeuroImage, 25*(4), 1068–1076. https://doi.org/10.1016/j.neuroimage.2004.12.050

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews. Neuroscience, 11*(8), 599–605. https://doi.org/10.1038/nrn2882

Kristop, C. A., Moreno, S. J., & Anta, J. F. (2020). What do listeners understand by "continuity" and "closure?" Tracking down the links between tonal expectancy, music training, and conceptualization. *Psychology of Music, 48*(3), 344–357. https://doi.org/10.1177/0305735618803000

Krumhansl, C. L., & Keil, F. C. (1982). Acquisition of the hierarchy of tonal functions in music. *Memory & Cognition, 10*(3), 243–251. https://doi.org/10.3758/BF03197636

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*(4), 334–368. https://doi.org/10.1037/0033-295X.89.4.334

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology. Human Perception and Performance, 5*(4), 579–594. https://doi.org/10.1037/0096-1523.5.4.579

Lamont, A. (1998). Music, education, and the development of pitch perception: The role of context, age, and musical experience. *Psychology of Music, 26*(1), 7–25. https://doi.org/10.1177/0305735698261003

Lamont, A., & Cross, I. (1994). Children's cognitive representations of musical pitch. *Music Perception, 12*(1), 27–55. https://doi.org/10.2307/40285754

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lin, H.-R. (2023, September 8–10). *Bedeutung von Geschlecht und sozioökonomischem Status für die Entwicklung musikalischer Fähigkeiten bei Jugendlichen: Vorläufige Ergebnisse der Langzeitstudie LongGold* [Relevance of gender and socioeconomic status for the development of musical skills in adolescence: Preliminary results of the longitudinal study LongGold] [Paper presentation]. The 39th annual meeting of the German Society for Music Psychology, Hannover.

Mack, M., Marie, D., Worschech, F., Krüger, T. H. C., Sinke, C., Altenmüller, E., James, C. E., & Kliegel, M. (2025). Effects of a 1-year piano intervention on cognitive flexibility in older adults. *Psychology and Aging, 40*(2), 218–235. https://doi.org/10.1037/pag0000871

Magne, C., Schon, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neuroscience, 18*(2), 199–211. https://doi.org/10.1162/jocn.2006.18.2.199

Maier-Karius, J., & Schwarzer, G. (2011). Die Beziehung zwischen Tonalitätsverstehen und kognitiven Fähigkeiten [The correlation between tonal understanding and cognitive abilities]. In W. Auhagen, C. Bullerjahn, & H. Höge (Eds.), *Jahrbuch der Deutschen Gesellschaft für Musikpsychologie* (Vol. 21, pp. 164–184). Hogrefe. https://doi.org/10.23668/psycharchives.2940

PsychOpen GOLD

Mandikal Vasuki, P. R., Sharma, M., Demuth, K., & Arciuli, J. (2016). Musicians' edge: A comparison of auditory processing, cognitive abilities, and statistical learning. *Hearing Research, 342*, 112–123. https://doi.org/10.1016/j.heares.2016.10.008

Matsunaga, R., Hartono, P., Yokosawa, K., & Abe, J. (2020). The development of sensitivity to tonality structure of music: Evidence from Japanese children raised in a simultaneous and unbalanced bi-musical environment. *Music Perception, 37*(3), 225–239. https://doi.org/10.1525/mp.2020.37.3.225

Miles, S. A., Miranda, R. A., & Ullman, M. T. (2016). Sex differences in music: A female advantage at recognizing familiar melodies. *Frontiers in Psychology, 7*, Article 278. https://doi.org/10.3389/fpsyg.2016.00278

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and Gestalt-like principles predict melodic expectations. *Cognition, 189*, 23–34. https://doi.org/10.1016/j.cognition.2018.12.015

Müllensiefen, D. (2022, September 2–4). *Die Entwicklung kognitiver und musikalischer Fähigkeiten: Vorläufige Ergebnisse einer Langzeitstudie* [The development of cognitive and musical abilities: Preliminary results from a long-term study] [Paper Presentation]. The 38th annual meeting of the German Society for Music Psychology, Würzburg. https://longgoldstudy.files.wordpress.com/2022/09/dgm2022_muellensiefen-et-al_vorlaeufige-ergebnisse.pdf

Müllensiefen, D., Elvers, P., & Frieler, K. (2022). Musical development during adolescence: Perceptual skills, cognitive resources, and musical training. *Annals of the New York Academy of Sciences, 1518*(1), 264–281. https://doi.org/10.1111/nyas.14911

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS One, 9*(2), Article e89642. https://doi.org/10.1371/journal.pone.0089642

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine, 21*(4), 125–127. https://doi.org/10.1136/ebmed-2016-110401

Paananen, P. (2003). *Monta polkua musiikkiin: Tonaalisen musiikin perusrakenteiden kehittyminen musiikin tuottamis- ja improvisaatiotehtävissä ikävuosina 6–11* [Many paths to music: The development of the basic structures of tonal music in music production and improvisation tasks in ages 6–11] [Unpublished doctoral dissertation, University of Jyväskylä, Finland]. http://search.ebscohost.com/login.aspx?direct=true&db=rih&AN=A534346&site=ehost-live

Paananen, P. (2007). Melodic improvisation at the age of 6–11 years: Development of pitch and rhythm. *Musicae Scientiae, 11*(1), 89–119. https://doi.org/10.1177/102986490701100104

Paananen, P. (2009). Children's and adolescents' harmonisation of a tonal melody. *Music Education Research, 11*(2), 153–174. https://doi.org/10.1080/14613800902923294

Patel, A. D. (2021, July 28–21). *Measuring and modeling melodic expectation using singing: Surprising findings and new opportunities for cross-cultural research* [Keynote]. ICMPC16-ESCOM11. Connectivity and diversity in music cognition, University of Sheffield, UK.

PsychOpen GOLD

Pinheiro, J., & Bates, D. (2024). *nlme: Linear and nonlinear mixed effects models* (Version 3.1–166) [Computer software]. https://CRAN.R-project.org/package=nlme

Piston, W. (1978). *Harmony* (4th ed.). Norton.

Platz, F., Kopiez, R., Lehmann, A. C., & Wolf, A. (2022). Measuring audiation or tonal memory? Evaluation of the discriminant validity of Edwin E. Gordon's "Advanced Measures of Music Audiation.". *Music & Science, 5*, Article 20592043221105270. https://doi.org/10.1177/20592043221105270

Politimou, N., Douglass-Kirk, P., Pearce, M., Stewart, L., & Franco, F. (2021). Melodic expectations in 5- and 6-year-old children. *Journal of Experimental Child Psychology, 203*, Article 105020. https://doi.org/10.1016/j.jecp.2020.105020

Prince, J. B., & Schmuckler, M. A. (2014). The tonal-metric hierarchy. *Music Perception, 31*(3), 254–270. https://doi.org/10.1525/mp.2014.31.3.254

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods, 10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Putkinen, V., Tervaniemi, M., Saarikivi, K., Ojala, P., & Huotilainen, M. (2014). Enhanced development of auditory change detection in musically trained school-aged children: A longitudinal event-related potential study. *Developmental Science, 17*(2), 282–297. https://doi.org/10.1111/desc.12109

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. https://www.R-project.org/

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift für Psychologie, 227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software, 93*(6), 1–51. https://doi.org/10.18637/jss.v093.i06

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Sauvé, S. A., Cho, A., & Zendel, B. R. (2021). Mapping tonal hierarchy in the brain. *Neuroscience, 465*, 187–202. https://doi.org/10.1016/j.neuroscience.2021.03.019

Schellenberg, E. G., Bigand, E., Poulin-Charronnat, B., Garnier, C., & Stevens, C. (2005). Children's implicit knowledge of harmony in western music. *Developmental Science, 8*(6), 551–566. https://doi.org/10.1111/j.1467-7687.2005.00447.x

Schmuckler, M. A. (2016). Tonality and contour in melodic processing. In S. Hallam, I. Cross, & M. H. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed., pp. 143–165). Oxford University Press.

Schwarzer, G., Siegiesmund, A., & Wilkening, F. (1993). Entwicklung des Tonalitätsverstehens bei der Beurteilung und Produktion von Liedschlüssen [Development of tonal understanding measured by ratings and productions of song endings]. In K.-E. Behne, G. Kleinen, & H. de La Motte-Haber (Eds.), *Jahrbuch der Deutschen Gesellschaft für Musikpsychologie. Empirische Forschungen—Ästhetische Experimente* (Vol. 10, pp. 75–90). Florian Noetzel.

PsychOpen GOLD

Serafine, M. L. (1988). *Music as cognition: The development of thought in sound*. Columbia University Press.

Speer, J. R., & Meeks, P. U. (1985). School children's perception of pitch in music. *Psychomusicology: Music, Mind, and Brain, 5*(1), 49–56. https://doi.org/10.1037/h0094200

Stalinski, S. M., & Schellenberg, E. G. (2010). Shifting perceptions: Developmental changes in judgments of melodic similarity. *Developmental Psychology, 46*(6), 1799–1803. https://doi.org/10.1037/a0020658

Swanwick, K., & Tillman, J. (1986). The sequence of musical development: A study of children's composition. *British Journal of Music Education, 3*(3), 305–339. https://doi.org/10.1017/S0265051700000814

Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity to western musical structure. *Journal of Experimental Psychology. Human Perception and Performance, 18*(2), 394–402. https://doi.org/10.1037/0096-1523.18.2.394

Trehub, S. E., Schellenberg, E. G., & Kamenetsky, S. B. (1999). Infants' and adults' perception of scale structure. *Journal of Experimental Psychology. Human Perception and Performance, 25*(4), 965–975. https://doi.org/10.1037/0096-1523.25.4.965

Turner, R. M., & Higgins, J. P. T. (2019). Bayesian meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 299–314). Russell Sage Foundation.

van Doorn , J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin and Review, 28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Model comparison and the principle of parsimony* (Vol. 1, pp. 300–319). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Verosky, N. J. (2021). Interpreting the tonal hierarchy through corpus analysis. *Psychomusicology: Music, Mind, and Brain, 31*(2), 96–106. https://doi.org/10.1037/pmu0000276

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3). https://doi.org/10.18637/jss.v036.i03

Virtala, P., Huotilainen, M., Putkinen, V., Makkonen, T., & Tervaniemi, M. (2012). Musical training facilitates the neural discrimination of major versus minor chords in 13-year-old children. *Psychophysiology, 49*(8), 1125–1132. https://doi.org/10.1111/j.1469-8986.2012.01386.x

von Helmholtz, H. (1896). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* [On the sensations of tone as a physiological basis for the theory of music] (5th ed.). Friedrich Vieweg und Sohn.

Vuvan, D. T. A. (2013). *The statistical learning of musical expectancy* [Doctoral dissertation]. University of Toronto (Canada).

PsychOpen GOLD

https://utoronto.scholaris.ca/server/api/core/bitstreams/fc3e21ef-4769-4fb1-9475-684fb229f7cf/content

Wehrum, S., Degé, F., Ott, U., Walter, B., Stippekohl, B., Kagerer, S., Schwarzer, G., Vaitl, D., & Stark, R. (2011). Can you hear a difference? Neuronal correlates of melodic deviance processing in children. *Brain Research, 1402*, 80–92. https://doi.org/10.1016/j.brainres.2011.05.057

What Works Clearinghouse. (2020). *What Works Clearinghouse™ procedures handbook, Version 4.1.* U.S. Department of Education, Institute of Education Science, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). *Bayesian meta-analysis with weakly informative prior distributions*. PsyArXiv. https://doi.org/10.31234/osf.io/7tbrm

Wilson, S. J., & Wales, R. J. (1995). An exploration of children's musical compositions. *Journal of Research in Music Education, 43*(2), 94–111. https://doi.org/10.2307/3345672

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*(6), 224–228. https://doi.org/10.1111/1467-8721.00154

Zhang, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2020). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music, 48*(3), 389–409. https://doi.org/10.1177/0305735618804038

Appendix

A1. Table

Table A1-1

Comparison of Statistical Models Based on Their Fit to the Data for Describing the Growth Trajectories Observed in Krumhansl and Keil (1982)

Model	AIC	BIC	RMSEA
Non-linear mixed model	6.61	10.60	0.03
Logistic model	39.97	43.96	0.09
Quadratic model	42.37	47.35	0.09
Cubic model	44.09	50.07	0.09
Linear model	47.07	51.05	0.09
Non-linear saturation model	48.94	52.92	0.12
Sigmoidal model	49.83	53.81	0.17

Note. AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; RMSEA: Root Mean Square Error of Approximation.

PsychOpen GOLD

A2. Figures

Figure A2-1

Posterior Probability Density for Modeled Parameters for the Bayesian Three-Level Model: (a) μ for the Mean, (b) τ_1 for Between-Study Standard Deviation, and (c) τ_2 for the Within-Study Standard Deviation

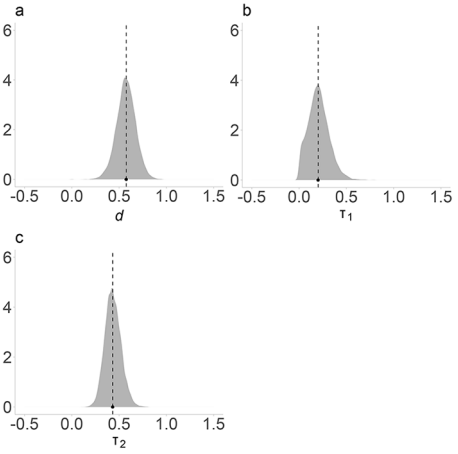
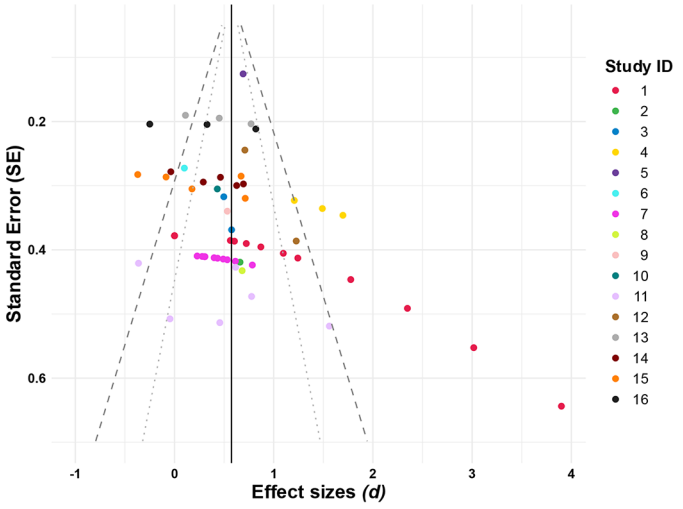


Figure A2-2

Funnel Plot of Reported Study Estimates by Standard Errors

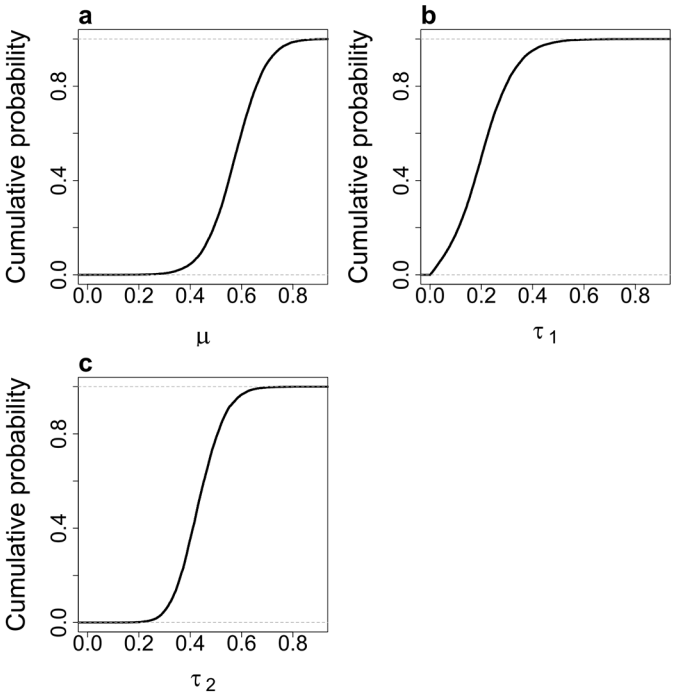


Note. Solid line: modelled mean effect; dashed lines (outer): 95% CI; dotted lines (inner): 80% CI. Each point represents an individual study, color-coded by Study ID.

PsychOpen GOLD

Figure A2-3

Empirical Cumulative Probability Function for Modeled Parameters for the Bayesian Three-Level Model: (a) $\boldsymbol{\mu}$ for the Mean, (b) $\boldsymbol{\tau}_1$ for Between-Study Standard Deviation, and (c) $\boldsymbol{\tau}_2$ for the Within-Study Standard Deviation



Jahrbuch Musikpsychologie (JBDGM) is the official journal of the German Society for Music Psychology (DGM).

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.
www.leibniz-psychology.org