Approaches: An Interdisciplinary Journal of Music Therapy
**16 (2) 2024**
ISSN: 2459-3338
https://doi.org/10.56883/aijmt.2024.64

ARTICLE

# The need for robust critique of arts and health research: An examination of the Goldbeck and Ellerkamp (2012) randomised controlled trial of music therapy for anxiety in children, and its treatment in four systematic reviews

**Stephen Clift**
Canterbury Christ Church University, UK

**Katarzyna Grebosz-Haring**
Paris Lodron University Salzburg &
University Mozarteum Salzburg, Austria

**Leonhard Thun-Hohenstein**
Paracelsus Medical University, Austria

**Anna Katharina Schuchter-Wiegand**
Paris Lodron University Salzburg, Austria

**Arne C. Bathke**
Paris Lodron University Salzburg, Austria

## ABSTRACT

We describe work-in-progress to conduct a systematic review of research on the effects of arts-based programmes for mental health in young people. We have searched for relevant studies through major databases and screened extant systematic reviews for additional research which meets our inclusion criteria. We have reservations, however, regarding both the quality of existing primary studies and of recently published systematic reviews in this area of arts and health. In a previous paper (Grebosz-Haring et al., 2022), we focused on a randomised controlled trial (RCT) on art therapy for adolescent girls with 'internalising' and 'externalising' problems, and its inclusion in three systematic reviews, and expressed concerns. In this paper, we extend the scope of our critical scrutiny to a research paper on music therapy with children described as having anxiety disorders (Goldbeck & Ellerkamp, 2012), and its treatment in four recent systematic reviews / meta-analyses (Ponomarenko et al., 2017; Cohen-Yatziv & Regev, 2019; Bosgraf et al., 2020). We demonstrate limitations in the Goldbeck and Ellerkamp study which undermine the conclusion they reach on the effectiveness of music therapy in the remission of anxiety disorders. We also show that the reviews are not sufficiently critical and make errors in the treatment of Goldbeck and Ellerkamp's research, which cast doubts on their dependability. Finally, we reflect on the lessons learned from our critique and draw some positive recommendations for future research and the conduct of reviews.

## KEYWORDS

music therapy,
children,
anxiety,
systematic review,
meta-analysis,
critique

**Publication history:**
Submitted 25 Mar 2022
Accepted 4 Jun 2022
First published 11 Aug 2022


**Publisher:** Queen Margaret University Edinburgh | Centre for Person-centred Practice Research

AUTHOR BIOGRAPHIES

**Stephen Clift** is Professor Emeritus, Canterbury Christ Church University, and former Director of the Sidney De Haan Research Centre for Arts and Health. He is a Professorial Fellow of the Royal Society for Public Health (RSPH) and is Visiting Professor in the International Centre for Community Music, York St John University, and the School of Music, University of Leeds. Since 2000 he has pursued research in arts and heath and particularly the potential value of group singing for health and wellbeing. Stephen was one of the founding editors of *Arts & Health: An International Journal for Research, Policy and Practice*. He is joint editor with Professor Paul Camic of the *Oxford Textbook of Creative Arts, Health and Wellbeing*. [stephen.clift@canterbury.ac.uk] **Katarzyna Grebosz-Haring** is a systematic musicologist, music educator and music therapist based in Salzburg, Austria. She is currently a senior scientist in the Inter-University Organization Science and Arts at the University of Salzburg and the Mozarteum University Salzburg. She has directed several empirical studies on the social and psychological meanings of music and art. Her main research interests are systematic-empirical approaches in music research, the clinical and educational application of music and art, and the mediation of music. She is a member of the Royal Society for Public Health. She has authored numerous interdisciplinary publications in SAGE, Routledge, Elsevier and others. [katarzyna.grebosz-haring@plus.ac.at] **Leonhard Thun-Hohenstein** obtained his medical degree in Vienna. He finished boards in *Pediatrics* (1986), *Child and Adolescent Psychiatry and Psychotherapeutic Medicine* (2018) and *Neuropsychiatry* (1993). He was former Head of Department and Professor for Child and Adolescent Psychiatry at the SALK, Campus CDK, Paracelsus Private Medical University in Salzburg. He is vice president of the Austrian Society for Child and Adolescent Psychiatry (ÖGKJP), board member of the Austrian Society for Child protection in medicine (ÖGKiM) and member of the Oberster Sanitätsrat (Supreme Medical Counsil, Federal Ministry for Health). [lthun@icloud.com] **Anna K. Schuchter-Wiegand** obtained her Magistra rer. nat. in Vienna (2015). Currently she is working as scientific staff at the Paris Lodron University Salzburg for the Project 'Art is a doctor: Research on the effect of musical activity on overall well-being of children and adolescents with mental illnesses and from socially underprivileged households.' [annakatharina.schuchter-wiegand@plus.ac.at] **Arne C. Bathke** obtained his PhD in Mathematics in Göttingen (2000), and he is Professor of Statistics at the University of Salzburg. His main methodological research areas are nonparametric statistics and inference for multivariate data. Currently, he is President of the International Biometric Society – Region Österreich-Schweiz (IBS-ROeS), on the board of the Austrian Statistical Association (ÖSG), Editor-in-Chief (with M. Schmid) of Biometrical Journal, as well as on the editorial board of two other international statistics journals (*International Journal of Biostatistics*, *Journal of the American Statistical Association*). [Arne.Bathke@plus.ac.at]

## INTRODUCTION

Mental health problems represent a major global concern among children and adolescents due to the high prevalence rate and their multifaceted nature. The development of mental disorders is complex not only because it involves multiple genetic and biological factors, but also because it involves psycho-social and behavioural risk factors (Grebosz-Haring & Thun-Hohenstein, 2018). Stressful experiences and chronic stress are above all relevant aetiological factors (Grebosz-Haring & Thun-Hohenstein, 2018). The challenge of mental disorders has led to the appearance of multimodal treatment concepts and new complementary therapeutic approaches that can attenuate stress, regulate emotions and enhance self-esteem, self-control, self-efficacy, spontaneity and creativity to improve everyday performance in social settings (Grebosz-Haring & Thun-Hohenstein, 2020). In this context, arts activities may provide a complement or alternative to biomedical and psychotherapeutic treatments. Based on emerging evidence, artistic activities such as musical activities can elicit positive feelings and influence hormonal system activity (stress response; Grebosz-Haring & Thun-Hohenstein, 2018; Grebosz-Haring et al., 2022). Furthermore, Grebosz-Haring & Thun-Hohenstein (2020) argue that engagement in arts activities can stimulate creative processes to increase conscious awareness and bring distraction, attention, imagery, joy, and pleasure. This can encourage young people to engage in a dialogue with themselves, other youth, their parents, and the wider social environment. These effects can be linked to mental health outcomes and can help with efforts to support or treat several mental health problems.

Grebosz-Haring and Thun-Hohenstein (2018) undertook a two-year pilot art and research project that ran in the University Clinic for Children and Adolescents Psychiatry at Christian Doppler Clinic / Paracelsus Medical University Salzburg. Young people experiencing mental health challenges had the

opportunity to engage in creative-artistic activities, including singing, music listening, textile design, drama, or clownery incorporated into traditional treatment routines to support creative expression. The preliminary results suggested that music and arts activities may provide benefits for young people with mental health problems. However, the authors identified major methodological challenges in setting up a controlled study with a larger group of young mental health patients in a clinical setting.

At this stage we decided that before designing a larger-scale trial, it was appropriate to conduct a systematic review of arts-based programmes for children and young people with a psychiatric diagnosis. Furthermore, we did not find a review that explored this issue in PROSPERO; the international prospective register of systematic reviews.

Several reviews have appeared recently that support the view that creative arts engagement can be beneficial for the health and wellbeing of children and young people. However, available systematic reviews (Glew et al., 2021; Mansfield et al., 2018) have not addressed our specific concern with the value of creative arts for children and young people with diagnosed mental health challenges, and other reviews are not systematic and insufficiently critical. Fancourt and Finn (2019), for example, report a scoping review of the arts and health research literature that includes diverse studies involving children and young people, but a critical perspective is lacking. Dowlen (2021) reports a rapid review of studies on creative arts and young people's mental health but excludes consideration of research on creative arts therapies. Clift et al. (2021) have argued that rather than scoping and rapid reviews, the field of arts and health "must rely on rigorous systematic reviews involving careful quality assessment of both quantitative and qualitative studies" (p.13).

We have, therefore, prepared a protocol for a systematic review of controlled studies of creative arts activities / arts therapy for children and young people experiencing mental health problems to appraise the quantitative evidence and synthesise established knowledge. The protocol (Grebosz-Haring et al., 2021) was developed in accordance with the latest PRISMA[1] guidelines (Page et al., 2021), and published through PROSPERO[2] (Page et al., 2018).

So far, we have searched major electronic databases, and supplemented this approach by cross-checking reference lists in relevant recent reviews. We have also used Google Scholar to identify citations of potentially relevant papers in subsequent publications. Our preparatory work, however, has revealed some concerns. Firstly, regarding the quality of published research on the effects of arts programmes and therapy for young people with mental health challenges, and secondly, a lack of criticality in recent reviews of this literature.

In a previous paper (Grebosz-Haring et al., 2022), we discussed a research paper by Bazargan and Pakdaman (2016)[3], which evaluated art therapy for adolescent girls identified as having 'internalising' or 'externalising' 'problems' and considered the treatment of this paper in three subsequent systematic reviews (Ponomarenko et al., 2017; Cohen-Yatziv & Regev, 2019; Bosgraf et al., 2020). We found substantial limitations in the design and execution of the Bazargan and Pakdaman (2016) research, and a lack of critical perspective in the three systematic reviews which included it.

[1] Preferred Reporting Items for Systematic Reviews and Meta-Analyses, https://prisma-statement.org/

[2] International Prospective Register of Systematic Reviews, https://www.crd.york.ac.uk/prospero/

[3] The research identified through our search was organised alphabetically by principal author, and the Bazargan and Pakdaman study was the first on our list.

Our critique therefore applied both to the original paper, and to weaknesses in the systematic reviews. In this paper, we repeat and extend this approach by considering one widely cited example of research on music therapy for children (Goldbeck & Ellerkamp, 2012) and the inclusion of this study in two systematic reviews (Belski et al., 2021; Ponomarenko et al., 2017), and two meta-analyses (Geipel et al., 2018; Lu et al., 2021). The Goldbeck and Ellerkamp (2012) study was chosen as it was the first music therapy paper on our alphabetically organised list of studies identified through the systematic search of databases.

The Goldbeck and Ellerkamp study is well designed, and clearly reported. The study was pre-registered with www.clinicaltrials.gov (NCT01062646), received ethical approval, included an estimation of required number of participants to be of sufficient power[4], and was conducted in accordance with CONSORT[5] guidance (Schulz et al., 2010). However, as we will see, it is not free from limitations and, perhaps more importantly, the treatment it receives in systematic reviews and meta-analyses is far from satisfactory.

As in our earlier paper, we set our discussion in the context of critical perspectives on the conduct of both systematic reviews and more especially meta-analyses in medicine, health care and education (Shamseer et al., 2015). Although both are considered as being at the top of most models of 'evidence hierarchies' – and meta-analyses have even been characterised as providing the 'platinum standard' in the synthesizing of evidence (Stegenga, 2011), substantial reservations have been expressed about the principles and practice of systematic reviews and meta-analysis and weaknesses in their execution.

MacLure (2005), for example, presents a detailed critique of the systematic reviews on educational topics, conducted, and supported, by the EPPI Centre at the University of London over the period 2002-4.[6] Greenhalgh et al. (2018) are critical of the view that systematic reviews are necessarily superior to much maligned narrative reviews. Ioannidis (2016) has been a particularly vocal critic of systematic reviews and meta-analyses, the production of which "has reached epidemic proportions" (p.487). He regards most systematic reviews and meta-analyses as "unnecessary, misleading, and/or conflicted" (p.468). Møller et al. (2018), go further and question whether systematic reviews and meta-analyses are a useful form of research, arguing that "many of them are focused on unimportant questions […] redundant and unnecessary", and "flawed beyond repair", with "only about 3% of them […] well done and clinically useful" (p.520).

Meta-analysis, is a form of systematic review in which the final step involves a statistical synthesis of quantitative findings from multiple sources. The procedure came in for early substantial criticism from Eysenck (1978), who referred to meta-analysis as "an exercise in mega-silliness" (p.517). His criticisms were elaborated in subsequent papers (Eysenck, 1984, 1994, 1995), which make trenchant points about the limitations of meta-analysis. Of these, the 'adding apples and oranges' problem is especially applicable to the critique of meta-analyses on music therapy and anxiety considered below:

[4] In the event, unfortunately, the target number was not achieved, and so the study was under-powered.

[5] Consolidated Standards of Reporting Trials, http://www.consort-statement.org/

[6] See: https://eppi.ioe.ac.uk/cms/ for current details of the work of the EPPI Centre.

> Meta-analysis is only properly applicable if the data summarised are homogeneous – that is, treatment, patients, and end points must be similar or at least comparable. Yet often there is no evidence of any degree of such homogeneity and plenty of evidence to the contrary. (Eysenck, 1994, p.791)

Eysenck (1978) was also critical of the inclusion of studies in early meta-analyses of variable methodological quality – what he refers to as the problem of "garbage in, garbage out" (p.517). Reservations have continued ever since, despite attempts to tackle these early criticisms (Sharpe, 1997). A stringent critique of meta-analysis comes from Stegenga (2011) who argues that meta-analysis is more subjective than generally claimed, given "the numerous decisions that must be made when designing and performing a meta-analysis" (p.505). We will demonstrate below the operation of such subjectivity in systematic reviews and meta-analyses which include the Goldberg and Ellerkamp study.

## THE GOLDBECK AND ELLERKAMP (2012) RCT ON MUSIC THERAPY FOR ANXIETY IN CHILDREN

Goldbeck and Ellerkamp (2012) report a 'pilot study' which investigates the 'efficacy' of 'Multimodal Music Therapy' (MMT)[7], for children with diagnosed anxiety disorders when compared to 'treatment as usual' (TAU). MMT is described as "a combination of music therapy and cognitive-behavioural therapy" (CBT) (p.395) and TAU was one of three forms of treatment available to the control group. Thirty-six children aged 8-12 years diagnosed by trained assessors as having an anxiety disorder were recruited to the study and randomly assigned to 15 sessions of MMT or to TAU. The programme also included three sessions for parents. Diagnostic status and dimensional outcome variables were assessed at the end of treatment and diagnostic status assessed again four months later.[8] MMT was found to be more effective compared to TAU according to the remission rates after treatment (MMT 67%; TAU 33%; $chi^2 = 4.0$; $p = 0.046$) and remissions persisted until four months post-treatment. Validated scales, including the State-Trait Anxiety Inventory (STAI-C), were completed by the children at baseline and after the intervention. In contrast to the clinical outcome, however, children showed equivalent improvement on several validated scales, including STAI-C, after both MMT and TAU. Goldbeck and Ellerkamp conclude that their results indicate that MMT is a 'promising' treatment for children with anxiety disorders. They recommend that further evaluation with larger samples and comparisons to 'pure CBT' are needed to further test their findings.

Goldbeck and Ellerkamp are commendably candid about the limitations of their study, which potentially compromise their conclusion regarding the effectiveness of MMT:

- Firstly, the study compared two treatments, MMT and TAU, with different degrees of standardization and different "dosage of application" (p.410). As a result, "the better response

[7] Multimodal Music Therapy is described in some detail in Table 1 of Goldbeck and Ellerkamp's report. They also state that a manual was created to guide music therapists in delivering the programme of activities. Unfortunately, the web-link provided no longer functions, and further searching failed to locate it. Sadly, Goldbeck died in 2017, and we have been unable to contact Ellerkamp for further information.

[8] The primary outcome variable is 'remission' of the anxiety disorder. Goldbeck and Ellerkamp explain why this is used to assess outcomes as follows: "remission of diagnosis is a central criterion for treatment response, as insurance companies pay treatment only indicated by diagnosis" (p.403). This is clearly specific to the German context.

rate in the MMT group might be due to non-specific general effects of child psychotherapy such as attention, dosage, or training of therapists, and not due to the specific interventions" (p.410).

- Secondly, although personnel undertaking post-treatment and follow up assessments were independent of the therapists, "not all evaluators were able to be blind to the intervention type. Thus, treatment expectancy of patients and of some evaluators may have influenced […] assessments" (p.410).
- Thirdly, the sample size was small, and the study was under-powered.
- Fourthly, "despite randomization, gender and subtypes of anxiety disorder were not equally distributed in both groups and therefore the better response rate in the MMT group may be due to the higher proportion of girls and of patients with social phobia" (p.410).
- And finally, the MMT programme was very multi-faceted, and included CBT methods, and so "the treatment effects might be more determined by the CBT modules than by the music intervention modules"[9] (p.410).

In addition, however, further critical points can be made, which go beyond the limitations they themselves acknowledge.

## A ROBUST CRITIQUE OF THE GOLDBECK AND ELLERKAMP STUDY

### The process of recruitment and the diagnosis of anxiety disorder

Goldbeck and Ellerkamp (2012) give the following account of how the children were identified for potential participation in the study:

> The study was announced in a local newspaper report and among community therapists. Children who responded to the newspaper announcement[10] or were referred to the study centre by community therapists or directly consulting the outpatient clinic of the Department of Child and Adolescent Psychiatry / Psychotherapy at the University of Ulm Medical Centre were screened for eligibility. (p.398)

Goldbeck and Ellerkamp are detailed in their account of the instrument used to establish 'diagnostic eligibility' using the 'KIDDIE-SADS' system (p.403), and they refer to the "gold standard for the assessment of mental disorders in children recommending structured clinical assessments integrating information from the child, a (parental) caregiver, and clinical judgement" (p.403). Making a diagnosis is one of the central duties of a medical doctor, for guaranteeing a standardised and evidence-based treatment. Thus, the procedure to diagnose anxiety disorder by a standardised

[9] It should be noted, however, that the character of the musical components appeared to have been guided by CBT principles, as indicated by the emphasis on 'relaxing' music.

[10] It is difficult to imagine that the children themselves responded to the newspaper announcement, and presumably their parents did so.

procedure is not per se a flaw – if the diagnostic process follows agreed scientific standards. However, their account of how the children were screened for an 'anxiety disorder' is very sparse, and no details are given on how information from children and parents was gathered and integrated with the clinical assessor's judgement:

> Sixty-two children were screened by telephone at the beginning of the study (see Figure 1). Fourteen of the screened potential participants were ineligible (e.g., no anxiety disorder), seven were not interested in participation, five refused for other reasons, such as nonavailability for regular treatment or assessments. Finally, 36 participants were included, completed the full baseline assessment, and were randomized to either MMT or TAU. (p.405)

The role of the parents in recruitment and screening stage is not specified, which is puzzling, given the age range of the children (8-12 years) and the fact that parents were active participants in the Multimodal Music Therapy programme. Surely, it would have been the parents who expressed interest in their child being part of the study in the first place, and the parents who would have refused participation on the grounds of 'nonavailability' given the conditions of involvement.

The age composition of the sample also deserves some comment. Children in Germany may transfer to secondary education from the age of ten, so some of the children may have been in secondary schools and others in primary. In addition, some of the girls may have already begun the transition into puberty. Neither of these issues is acknowledged or discussed by Goldbeck and Ellerkamp but they may well have a bearing on the children's engagement with music therapy, especially in a group setting.

## The use of standardised, validated scales in assessment

In addition to the 'clinical' interviews which provided the primary diagnosis of an 'anxiety disorder,' Goldbeck and Ellerkamp (2012) also employed several standardised and validated scales which purport to measure a range of psychological constructs: state and trait anxiety, depression, social phobia, complaints, quality of life, and well-being (pp.403-404). However, no normative data for these scales are given; no cut-off points for 'clinical significance,' and no estimates of 'minimal clinically important difference' (MCID) scores. Consequently, the mean values on these measures, reported in Table 3, are difficult to interpret, without further inquiry into their psychometric properties. It is also widely reported that the prevalence of anxiety problems is greater in girls and women (Strand et al., 2021), but the paper does not acknowledge or discuss the implications of this difference. One measure of particular interest, given the clinical diagnosis of anxiety, is the children's version of the Spielberger State-Trait Anxiety Inventory (STAI-C). In the Goldbeck and Ellerkamp study, the 'trait' scale is used.[11] This consists of 20 statements, with a three-point Likert scale (1-3) giving a total score ranging from 20-60 and a mid-point of 40. The higher the score the higher the degree of anxiety. Table 3 in Goldbeck and Ellerkamp shows that the children in the study in the MMT group at baseline had a mean score of 48.1, and the control group had a mean score of 51.4. These values represent an average item score

[11] As we will see below, Geipel et al. (2018) and Lu et al. (2020) use the data from the STAI-C in their meta-analyses.

of approximately 2.5 and clearly indicate that the children were reporting high levels of anxiety. For both groups, scores were significantly lower at post-test (main effect time p=0.003), but still relatively high at 42.6 for the MMT group and 44.7 for the TAU group. However, Goldbeck and Ellerkamp do not report changes in STAI-C scores for children who are said to show remission, so it is difficult to judge whether this reduction in trait anxiety is clinically meaningful. The effect size for the change on the STAI-C for children in the MMT condition is estimated as 0.34 from the data reported in Table 3. This is half the value for the average effect size for therapeutic 'treatment gain' on the STAI-C trait scale reported in a meta-analysis of seven studies (Seligman et al., 2004).

In our view, the picture that emerges from the data in Table 3 indicate that both MMT and TAU groups substantially improved on the Child Behaviour Checklist scales, and the State Trait Anxiety Inventory trait measure.

## The nature and appropriateness of Multimodal Music Therapy (MMT)

It is also necessary to question the nature of the MMT programme (described in detail in their Table 1), and the lack of rationale for this approach to treating children with anxiety disorders. From a behavioural therapy standpoint, if a child is diagnosed as having a specific, disabling phobia, or specific form of severe anxiety, surely the treatment approach should be a carefully planned and individually tailored programme of behaviour therapy. The stated justification for a musical element is that music allows children to express themselves non-verbally, when talking about the challenges they are facing may be difficult.

As the treatment was not delivered individually, apart from the first three sessions, MMT appears to be as a generic treatment not specifically tailored to the challenges facing individual children. This may have happened initially in the three individual sessions, but it is not clear how it would have happened in the nine group sessions involving 18 children, with a wide age range (from 8-12 years).

It is possible that the treatment process would itself be a potential trigger for anxiety in at least some of the children. If a child, for example, has a 'general anxiety disorder' or 'separation anxiety' would a new experience of engaging in therapy not raise their anxiety levels? Similarly, if a child has a general 'social phobia' or a fear of open spaces, might the new experience of therapy raise their fears? In which case, perhaps MMT, with its many components worked through a general process of de-sensitisation? The repeated references to 'relaxation' in the description of the programme points in this direction (the word 'relaxation' is used 17 times). It could be, in other words, that the whole programme provided a general 'counter-conditioning' experience for the children (Keller et al., 2020). Certainly, we can assume that the professionals delivering the MMT programme would have done their utmost to ensure that the experience was non-threatening and enjoyable for the children.

## The non-standardisation of treatment as usual (TAU)

There are also concerns about the notion of TAU. Firstly, typically in controlled trials of a new intervention, TAU should be available to both the intervention and control group, as it would be unethical to withhold accepted standard treatments for a diagnosed condition from the experimental group. Secondly, it is not clear that the children in the trial were existing patients of the psychiatric

service and so in receipt of treatment, as participants were 'recruited' in a variety of ways, including by responding to advertising. And thirdly, the notion of TAU is somewhat vague, given that there are three forms of treatment specified (brief behavioural interventions, psychodynamic psychotherapy, nonspecific group therapy).These were of varied duration, with no evidence presented that they were all considered to be equivalent evidence-based options. In addition, some children in the TAU group did not receive any treatment at all and were on a waiting list for the duration of the trial.

## Intention to treat vs. per protocol analysis of results

Goldbeck and Ellerkamp (2012) undertake an 'intention to treat' (ITT) analysis for their primary outcome measure of remission of anxiety disorder. Following treatment 12 out of 18 of the MMT group were judged to have improved clinically, as compared with 6 out of 18 children assigned to TAU. This difference is significant (just) at the 5% level (p=0.046). At four months follow up, the respective figures continue to be 12 out of 18 and 6 out of 18 but further attrition had occurred. Goldbeck and Ellerkamp do not report the result of a chi-square test for these data as they remain unchanged. However, the picture looks very different if a 'per protocol' (PP) analysis is undertaken. Given attrition of the sample, following treatment, the figures for MMT are 12 out of 16, and for TAU 6 out of 10, a difference which is not significant (p=0.42). At 4-month follow up, the values for MMT are 12 out of 16, and for TAU 6 out of 9, a difference which again is not significant (p=0.66).

In relation to the data gathered using validated scales, the picture is somewhat unclear. It might be expected that Goldbeck and Ellerkamp would follow the logic of ITT in analysing these data comparing baseline and post-treatment assessments (there was no four-month follow up with the validated scales). Reference to their Table 3, however, shows that degrees of freedom (df) reported for each of the scales vary, and no explanation is offered in the text. A footnote in the table refers to the fact that some families declined the intervention (for TAU that is) post-randomisation, and that some declined follow-up (in both arms of the trial), which implies that these factors affected the analysis. Another factor that might well account for varying df values is missing data for some individuals, but Goldbeck and Ellerkamp do not mention missing data in the text. However, none of these considerations is relevant to an ITT analysis, as baseline values would be employed at follow up and the df values would be consistent at 34 (N-k).

There are strong justifications for an ITT analysis, and limitations associated with PP analyses (Ranganathan et al., 2016; Tripepi et al., 2020). Ranganathan et al. (2016) point out that compared with an ITT, a PP analysis may exaggerate treatment effects, but note that both forms of analysis are recommended by the 2010 CONSORT guidelines (Schulz et al., 2010), so that the reader can more fully interpret the findings from a trial. Tripepi et al. (2020) present a balanced account of the pros and cons of ITT and PP analyses, pointing out that ITT analysis assesses the effect of 'assigning' a treatment (which may not be received), whereas PP analysis measures the effect of 'receiving' the treatment. In their view both approaches "are essentially valid but they have different scopes and interpretations dependent on the context" (p.513). What this all comes down to, is the question of potential biases associated with PP analyses and associated risks.

In the Goldbeck and Ellerkamp study, the randomisation process did not generate equivalent groups on at least two important factors (sex and diagnostic category), and there are other important sources of bias at work (not least non-blinding of some of the assessments of anxiety at post-test and follow-up). Also, the PP analysis reported above clearly does not support the finding of MMT superiority over TAU, which emerges from the ITT analysis.

A further consideration is that the p-value associated with the ITT analysis of remission data is just under the 0.05 critical value for significance. A more cautious approach, given the small size of the study, and the potential biases involved in the study, would be a more stringent statistical criterion for testing whether MMT leads to 'remission' of anxiety. This would a sensible approach, given also that the validated scales employed provided no evidence of greater benefit from MMT, compared with TAU. The use of a more stringent p-value for significance is also in line with recent recommendations that the p=0.05 criterion should generally be replaced with a value of p=0.005 (Benjamin et al., 2018).

## TREATMENT OF THE GOLDBECK AND ELLERKAMP RCT IN SYSTEMATIC REVIEWS / META-ANALYSES

Two systematic reviews and two systematic reviews plus meta-analyses include the Goldbeck and Ellerkamp (2012) study. An overview of the four reviews is provided in Table 1, following the criteria offered by the AMSTAR-2 rating system for assessing the quality of systematic reviews and meta-analyses (Shea et al., 2017). The Goldbeck and Ellerkamp study is the only research study concerned with music therapy in the treatment of children with anxiety problems included in these reviews. The following sections consider the account each review gives of this study. We will then comment on differing results using the Cochrane 'Risk of Bias' tools (Higgins et al., 2011; Sterne et al., 2019) in three reviews, and problems associated with the quantitative syntheses reported in the two meta-analyses.

### Ponomarenko et al. (2017): Investigating the efficacy of art and music therapy with vulnerable children and young people

Ponomarenko et al. (2017) provide an accurate account of the Goldbeck and Ellerkamp study and offer some critical comments. In the main these reflect the limitations that Goldbeck and Ellerkamp themselves acknowledge, but they offer the following insightful comment regarding the lack of correspondence between the change seen on the primary outcome measure (the clinical assessment of 'remission'), and the lack of difference in change between MMT and TAU on the standardised measures:

> […] although the principal measure, the KIDDIE-SADS tool, showed divergence between the experimental and control group, self-reports and parental measures did not identify change between the two groups. Whilst this does not necessarily indicate fallibility of the primary measure, it is interesting to note this difference and it raises questions about how 'improvement' is measured and categorised if it is not recognised by the participant and/or their parents. (pp.55-56)

| AMSTAR-2 questions | Pono-marenko et al. (2017) | Geipel et al. (2018) | Lu et al. (2021) | Belski et al. (2021 |
|---|---|---|---|---|
| 1 Did the research questions and inclusion criteria for the review include the components of PICO (i.e., population, intervention, control, and outcomes). | Partial yes | Yes | Yes | Yes |
| 2* Did the report of the review contain an explicit statement that review methods were established prior to the conduct of the review? Did the report justify any significant deviations from the protocol? | No | No | No | No |
| 3 Did the review authors explain their selection of study designs for inclusion in the review? | No | Partial yes | No | Yes |
| 4* Did the review authors use a comprehensive literature search strategy? | Partial yes | Partial yes | Partial yes | Yes |
| 5 Did the review authors perform study selection in duplicate | No | Yes | Yes | Yes |
| 6 Did the review authors perform data extraction in duplicate? | No | No | Yes | Yes |
| 7* Did the reviewers provide a list of excluded studies and justify exclusions? | Partial yes | Yes | Partial yes | Partial yes |
| 8 Did the authors describe the included studies in adequate detail? | Yes | Yes | Yes | Yes |
| 9* Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies in the review? | Partial yes | Yes | Yes | Yes |
| 10 Did the review authors report on the sources of funding for the studies included in the review?[12] | No | No | No | No |
| 11* If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results? | N/A | Yes | Yes | N/A |
| 12 If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis? | N/A | No | No | N/A |
| 13* Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review? | N/A | Yes | Yes | N/A |
| 14 Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity in the results of the review? | N/A | Yes | Yes | N/A |
| Q15* If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review? | N/A | Yes | Yes | N/A |
| Q16 Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? | Partial yes | Partial yes | Yes | Yes |

**Table 1:** AMSTAR-2 assessment of four systematic reviews of music therapy research
**Key:** * Seven 'critical' items are identified by Shea et al. (2017)

[12] The issue of funding is of relevance in trials evaluating drug treatments (which may be sponsored by industry or by independent agencies) but is not relevant to trials of music therapy and other psychological treatments where there is no commercial interest.

There are, however, several misunderstandings in Ponomarenko et al.'s comments that are worth unpacking. Firstly, both the clinical assessments and the standardised scales showed positive changes for both groups over time, but what is different is that the change in clinical status is significantly greater for the MMT group than the control group, whereas no significant interaction term emerges for any of the scales. Secondly, Goldbeck and Ellerkamp report no analysis that allow the reader to judge whether measures were consistent or not. For example, no information is given on whether the children who showed remission, also showed a significant reduction in self-assessed anxiety or other measures. Thirdly, Ponomarenko et al. make no mention of the fact that the remission data are subject to an ITT analysis, whereas the standardised measures appear to be analysed by a PP analysis. As noted above, a PP analysis of the remission data shows no greater benefit from MMT.

## Geipel et al. (2018): Music-based interventions to reduce internalising symptoms in children and adolescents

Geipel et al. (2018) report a meta-analysis of 'music-based interventions' to reduce 'internalising symptoms' in children and adolescents. A clearly documented process of searching, selection of relevant reports according to inclusion criteria, and screening of full text papers, results in five research reports for the meta-analysis, including Goldbeck and Ellerkamp. In what is the most informative and interesting section of their paper, Geipel et al. provide a traditional narrative review of the five studies (six paragraphs on p.652).[13] This is what Geipel et al. have to say about the Goldbeck and Ellerkamp study:

> In a randomized controlled trial Goldbeck and Ellerkamp (2012) investigated the efficacy of a joined music therapy and CBT program compared to treatment as usual in 36 children with a mean age of 9.94 years, who endorsed diverse anxiety disorders. Patients received three single sessions of 60 min each, nine group sessions of 100 min each and three group sessions of parent training. Mean duration of the program was 17.6 weeks. The program combined music therapy techniques as free and structured improvisation, dialogue music playing, musical expression of emotions, receptive music therapy methods for relaxation and cognitive-behavioural interventions as psychoeducation, social skills training, exposure to anxiety evoking stimuli and other creative techniques as therapeutic drawing. The primary outcome was the presence of an anxiety disorder measured by the Schedule for Affective Disorders and Schizophrenia for School-Age Children – Present and Lifetime Version (KIDDIE-SADS) (Kaufman et al., 1996). According to the reported remission rate, music therapy was superior to treatment as usual. Both groups showed a significant reduction in the STAI-C T-value, but no significant main effects of group assignment or a significant interaction effect of group assignment and time of measurement occurred. (p.652)

[13] See Greenhalgh et al. (2018) for a discussion of the respective merits of narrative and systematic reviews.

This is reasonably accurate as a summary except that:

- While 36 children were randomised (18 in each group), fewer children remained in the study at the end of treatment due to attrition (16 intervention and 14 control) and further children were lost at 4-month follow-up.
- They correctly state that music therapy was "superior to treatment as usual in terms of remission rate" (p.652) and state that this was Goldbeck and Ellerkamp's primary outcome measure, but they fail to discuss why they chose to ignore this measure and focus instead on the results from the STAI-C on 'trait anxiety' for inclusion in their meta-analysis.
- Interestingly, as they note, both the intervention and control groups showed significant changes on the STAI-C, but no significant interaction effect was reported. In the presentation of the results of the meta-analysis (Figure 2, p.651), the 'forest plot' of standardised mean differences, relates to the difference in MMT and TAU at follow-up, and this places the Goldbeck and Ellerkamp result in the lowest ranked position.

Finally, Geipel et al. (2018) neglect to mention the other secondary outcomes, in particular the scale used to assess depression. Scores from the Children's Depression Inventory might have been more appropriately included in the meta-analysis, given that the scores employed in the other four studies were from measures of depression.[14]

Elsewhere in the text, Geipel et al. note that the Goldbeck and Ellerkamp study was the only research to include participants with anxiety disorders – hence it was included because the review was broadened to cover 'internalising' symptoms, rather than having a narrower focus on depression. As with the other four studies included in the meta-analysis the Goldbeck and Ellerkamp study was judged to have a 'high risk of bias' on account of lacking 'blinding of participants and personnel' and lack of 'blinding of outcome assessment' (p.651). They also reiterate, in discussing the wide diversity of 'music therapeutic interventions' across the five studies, that Goldbeck and Ellerkamp "tested a multimodal therapy program adding adjuvant parent training" (p.653). They then make the following general comment about the five studies:

> Within these designs, it is impossible to distinguish which elements of the program were particularly helpful for the patients. Further music therapy frequently adopts a psychotherapeutic (often CBT) approach. Therefore, music therapy cannot be understood as [a] unique treatment approach but comprises distinct techniques to deliver psychotherapeutic (i.e., CBT) treatment and content. (p.653)

The main problem with Geipel et al.'s use of the Goldbeck and Ellerkamp study in their meta-analysis, is the fact that they ignore the principal outcome variable (remission), and instead utilise the post-test results from the STAI-C scale for the children in the MMT and TAU groups. This is inappropriate, as the key issue is the relative changes for the experimental and control groups between

[14] Had they done this, the title of Geipel et al.'s paper could have referred to 'symptoms of depression' rather than 'internalising symptoms.'

pre and post-test reflected in the interaction term in the repeat measures analysis of variance reported by Goldbeck and Ellerkamp.

## Lu et al. (2021): Effects of music therapy on anxiety: A meta-analysis of randomised controlled trials

Lu et al. (2021) report a wide-ranging review and meta-analysis of 32 controlled studies of the effectiveness of music therapy in addressing anxiety issues with diverse populations in different settings. Goldbeck and Ellerkamp (2012) is the only research paper involving children. As with the Geipel et al. meta-analysis, rather than focusing on the primary outcome in the Goldbeck and Ellerkamp evaluation (remission anxiety) they chose to focus on one of the secondary outcomes – scores on the STAI-C – and do so without explanation or justification.

Of more concern, however, is the fact that in their Figure 3 (p.7), they report the baseline results for the intervention and control groups on the STAI-C, to indicate the effect of MMT vs TAU. This is entirely incorrect and, moreover, the sample sizes cited are wrong. Then, in their Figure 4 (p.7), they incorrectly present the post-intervention results for the STAI-C as the results from a 4-month follow-up. Goldbeck and Ellerkamp did follow up 4-months after the end of the intervention, with an assessment of continued 'remission,' but not with the standardised scales. These errors committed by Lu et al., are particularly unfortunate as they report that data extraction was undertaken by two members of the review team independently (see Table 1).[15]

## Belski et al. (2021): The effectiveness of musical therapy in improving depression and anxiety among children and adolescents

Belski et al. (2021) report a systematic review of randomised controlled trials assessing the effectiveness of music therapy for treating anxiety and depression in children and adolescents. The review involves a qualitative synthesis and does not attempt a meta-analysis, as this was considered inappropriate due to "considerable clinical heterogeneity" (p.3) across the studies included. The scope of the review is similar to Geipel et al. (2018), but only three studies are common to the two reviews – one of which is Goldbeck and Ellerkamp. As Belski et al. is a later date, it includes three studies that were published after the period covered by the Geipel et al. review. Belski et al. employ the current Cochrane Risk of Bias Tool (RoB2) (Sterne et al., 2019) to assess the trials included, as opposed to the first version (Higgins et al., 2011) used by Geipel et al. and Lu et al.

Unfortunately, there are some errors in the Belski et al. review in their treatment of the Goldbeck and Ellerkamp study:

- They characterise Multimodal Music Therapy correctly as "active and receptive" but say that it "did not utilize a theoretical approach" (p.5). This is puzzling as Goldbeck and Ellerkamp clearly describe their model as a combination of cognitive-behavioural therapy and music

[15] The errors noted only came to light because the starting point for the exercise in this report was the Goldbeck and Ellerkamp study, but it seriously calls into question the accuracy of the entire meta-analysis reported by Lu et al. The larger challenge raised here is that this error was not identified during peer review of the Lu et al. paper prior to publication in the Elsevier journal Psychiatry Research.

and refer repeatedly to the key theoretical mechanism of relaxation. In addition, Belski et al. make no mention of the active role of parents in the therapy programme.

- They ignore the primary outcome measure of remission of anxiety as assessed by a clinician, and instead focus on the non-significant findings from standardised scales for depression and anxiety. In this respect, they misrepresent the outcome of the Goldbeck and Ellerkamp study.
- They refer to the use of intention to treat analysis, but mistakenly imply that this approach was applied to the secondary outcome measures, whereas it is clear from the CONSORT diagram reported by Goldbeck and Ellerkamp (p.406), and from the degrees of freedom values reported in Table 3 (p.409), that a per protocol analysis was performed on the standardised scale results.
- In reporting the scale scores for depression and anxiety at post-test, Belski et al. indicate that the sample sizes for the intervention and control groups were 16 in each case. This is an error as the CONSORT diagram and Table 3 show clearly that attrition occurred in both groups over the course of the trial. Nor do they refer to the fact that in analysing the scale data Goldbeck and Ellerkamp correctly used repeat measures ANOVA.
- Finally, Belski et al. suggest that the follow up period for assessment using the standardised scales was 16 weeks, but in fact this assessment took place immediately after the therapy programme; it was a further clinical assessment of remission that occurred after four months.

## FURTHER CRITICAL REFLECTIONS ON THE SYSTEMATIC REVIEWS AND META-ANALYSES

Table 1. above, presents a profile of each of the reviews using the AMSTAR-2 instrument[16] (Shea et al., 2017). All four reviews emerge as satisfactory in terms of the AMSTAR-2 criteria, although as we have seen the fact that two members of a review team were involved in independent selection of trials, or the extraction of details and data, does not guarantee that their judgements are accurate (Stegenga, 2011).

There are three further critical reflections on the reviews presented in this section of our paper. Firstly, although Geipel et al., Lu et al., and Belski et al., undertake 'risk of bias' assessments of the trials they include, significant concerns over subjectivity emerge when a comparison is made of these assessments for the Goldbeck and Ellerkamp study. Secondly, while the meta-analyses undertaken by Geipel et al. and Lu et al. appear to follow standard procedures and are reported fully, there is reason to doubt the legitimacy of pursuing quantitative synthesis given the heterogeneity of the studies. And thirdly, it is important to ask whether we learn anything important about the therapeutic value of music from the reviews, over and above the individual studies.

[16] A MeaSurement Tool to Assess systematic Reviews, https://amstar.ca/Amstar-2.php

## Subjectivity in using the Cochrane Risk of Bias tools

Geipel et al. and Lu et al. employ the first version of the Cochrane Risk of Bias Tool (RoB) (Higgins et al., 2011) in assessing the trials in their reviews, whereas Belski et al. make use of the second, revised version of this tool (RoB2) (Sterne et al., 2019). The two versions cover essentially the same threats to the validity of trials (such as problems with the randomisation process), and there is no space here to consider the precise details of the changes between the initial and revised tools. It is sufficient for our purposes to present the combined risk of bias assessments for the Goldbeck and Ellerkamp trial in Table 2. This shows that there is no consistency across the three reviews in the judgements made on randomisation, blinding of participants and personnel, blinding of outcome assessment and selective reporting. The failure to agree with respect to blinding of participants is especially surprising as it is obvious that the children and their parents were aware of their allocation to music therapy or TAU.

The only criterion on which the three teams agree is that there was low risk of bias due to lack of outcome data. In the initial version of the RoB tool, this source of bias is referred to as 'attrition bias.' In RoB2 tool, however, the phrase 'attrition bias' is abandoned and the guidance for this criterion is: "Were the data that produced this result analyzed in accordance with a prespecified analysis plan" (p.4). On this basis, the three review teams made an accurate judgement as the primary outcome measure of remission was subject to an ITT analysis, and the CONSORT flowchart reported by Goldbeck and Ellerkamp indicates that all participants initially randomised were included in the analysis. However, the CONSORT diagram also shows clearly that there was attrition, and this attrition clearly affected the data gathered from the structured questionnaires employed by Goldbeck and Ellerkamp (see Table 3, p.409). For the scale outcome data, therefore, all teams have made erroneous judgements.

| RoB Criteria (Higgins et al., 2011) used by Geipel et al. and Lu et al. | Geipel et al., 2018 | Lu et al., 2021 | Belski et al., 2021 | RoB2 Criteria (Sterne et al., 2019) used by Belski et al. |
|---|---|---|---|---|
| Random sequence generation (selection bias) | | | | Bias arising from the randomisation process |
| Allocation concealment (selection bias) | | | N/A | Bias arising from period and carryover effects |
| Blinding of participants and personnel (performance bias) | | | | Bias due to deviations from intended intervention |
| Blinding of outcome assessment (detection bias) | | | | Bias in measurement of the outcome |
| Incomplete outcome data (attrition bias) | | | | Bias due to missing outcome data |
| Selective reporting (reporting bias) | | | | Bias in selection of the reported result |
| Other bias | | | N/A | Other bias |
| Overall risk of bias | | | | Overall risk of bias |

**Table 2:** Risk of Bias assessments of Goldbeck and Ellerkamp (2012) in three systematic reviews
**Key:** green = low risk of bias, yellow = unclear, red = high risk of bias

## The problem of 'apples and oranges' in meta-analysis

The main problem with the two meta-analyses, is that the authors proceeded with a quantitative synthesis, when the heterogeneity of the studies indicates, as Belski et al. acknowledge, that such an exercise is inappropriate.

In Geipel et al. the final five studies included are very diverse and have little in common:

- Each study was conducted in a different country (Australia, Germany, South Korea, Taiwan, and United States)
- The ages of the participants vary (with one study including adults, notwithstanding the title of their review)
- The character of the interventions is very different (music medicine, music therapy and music education), and
- The outcome measures included are different (four assess depression each with a different measure and one assesses anxiety)

The problem of 'heterogeneity' comes to fore at a late stage in the meta-analysis as one of these studies is dropped following examination of the funnel plot (it is concerned with 'music medicine'). Finally, of the remaining four studies, only two reported significant positive outcomes from the intervention evaluated, with the other two studies showing no benefit from music therapy compared with the control (one of which is the Goldbeck and Ellerkamp paper).

The issue of diversity in the studies included in the Lu et al. (2020) meta-analysis is even more marked and is so wide-ranging, that it represents a textbook case of the 'apples and oranges' problem (Sharpe, 1997; Sharpe & Poets, 2020). Table 2 in the Lu et al. paper shows the diversity very clearly:

- Studies from 10 different countries (11 United States, 7 China, 2 each from Norway, Finland, Iran, Italy, and Brazil, and 1 from Germany, France, and Greece)
- Widely diverse population groups (e.g., Mexican farmworkers in their 30's living in the USA, institutionalised adults in their 80's with dementia in China, and patients aged 18-50, with obsessive compulsive disorder in Iran)
- Variations in the character, timing, and delivery of the 'music-based' intervention (i.e., active, passive and a combination, delivered individually or in groups), and finally,
- Variations in measured outcomes (no fewer than 15 different measures of anxiety).

Both meta-analyses rest upon a reification of 'internalised' problems and 'anxiety' – in other words an assumption that 'depression' and 'anxiety' exist as tangible 'things,' irrespective of an individual's culture, social circumstances, personal history and method of assessment. In the Lu et al. study, this means that the situational 'anxiety' experienced by a cardiac patient about to undergo an operation is the same as the long-term anxiety of a child diagnosed with a 'psychiatric disorder;' and that the anxiety of mothers with preterm babies, is the same as the anxiety experienced by male prisoners languishing in a Chinese prison. Equally, it is assumed that all the measuring instruments employed in the various studies are reliable, valid, and 'sensitive to change' and thus interchangeable

in measuring the same 'thing.' Given the diversity in the studies included, it is doubtful that the assessments of anxiety can be combined into a single meaningful estimate of effect size.

### What do the reviews add to an understanding of the therapeutic powers of music?

Sadly, we learn nothing new from the reviews about the therapeutic value of music. All we are given are vague generalisations that music can provide a 'distraction' from worries or can be an aid to 'relaxation' (Lu et al., 2021, p.8). These are experiences that most of us will have had at some point in our lives, and amount to little more than 'common sense.' There is mention by Lu et al. of the ways in which therapists have at their disposal aspects of music, "such as melody, timbre, rhythm, harmony, and pitch, to support and enhance physical, psychological and social well-being" (p.2), but nowhere in the review is there discussion of how these different components of music might contribute to therapeutic benefits.

## CONCLUSIONS, LIMITATIONS, AND RECOMMENDATIONS

### Conclusions

In this paper we have taken a target paper by Goldbeck and Ellerkamp (2012) evaluating MMT for children with diagnosed anxiety disorders and have considered the way in which this paper is treated in four systematic reviews, two of which conduct a meta-analysis. We have undertaken a robust critique of the initial study, and of the reviews, and the reader may feel that our analysis is negative and lacks balance. In conclusion, therefore, we offer some constructive reflections and positive recommendations.

Notwithstanding the critical issues we have raised in relation to the Goldbeck and Ellerkamp study, we have stated several times that the trial was well designed, conducted, and reported, and conforms to current standards with respect to pre-registration, the use of the CONSORT framework, ethical review, a detailed description of the music therapy programme (Robb et al., 2018), and attention to the issue of statistical power. No study is ever free from limitations, but the Goldbeck and Ellerkamp study was innovative and important, and it is for music therapy researchers to consider why such a significant study has never been replicated.

We accept that there is a role for systematic reviews and meta-analyses of RCTs, where the starting point is a question formulated in terms of a specific population, a clearly defined intervention, the use of relevant control conditions, and common or equivalent outcomes (e.g. the PICO formula). If an existing corpus of research is highly varied in these respects, then a research mapping or scoping review might be worthwhile, but not a systematic review or meta-analysis. For example, we might, based on the Goldbeck and Ellerkamp study, consider a review of research on MMT, with children diagnosed with anxiety disorders, where the control is 'TAI' and the outcome is 'remission' of anxiety. What we would find, however, is that only the study by Goldbeck and Ellerkamp meets the inclusion criteria. If there were at least several such studies, then a systematic review and even a meta-analysis would be worthwhile.

Our critique raises questions about the conduct and reporting of systematic reviews and meta-analyses, and processes of peer review which lead to such reviews to be published. As we have noted the reviews we consider appear to have been undertaken systematically according to widely accepted standards (as judged by applying the AMSTAR2 tool), and of course they have been published in peer-reviewed journals. It is only when we look carefully at the details of how a study they include in common, is treated, that problems appear.

## Limitations

There are limitations to the work we report here. We have only undertaken an analysis of one target paper by Goldbeck and Ellerkamp (2012) and considered the way it is treated in four systematic reviews/meta-analyses. There is no basis in what we report here for generalising beyond the papers we have considered. However, an earlier paper (Grebosz-Haring et al., 2022) showed that the findings from a RCT of art therapy were taken at face value in subsequent systematic reviews despite substantial limitations in the target study. We are repeating our approach in a critique of a controlled trial on dance-movement therapy and its treatment in nine evidence reviews. Our conclusion will again be that findings are taken at face value in the reviews with little acknowledgement of serious limitations in the target study.

## Recommendations

- Further studies following the innovative method demonstrated in this paper are needed to assess the accuracy and credibility of systematic reviews in the field of arts and health.
- Systematic reviews should be properly focused, pre-registered in PROSPERO (Page et al., 2018) and conducted to a high standard following current PRISMA guidelines (Page et al., 2021). Particular attention to double checking judgements of bias and ensuring accuracy in the process of data extraction.
- Peer review of reports of systematic reviews and meta-analyses needs to be rigorous and involve careful checking of the accuracy of how primary sources are treated.
- Greater attention is needed in the field of arts and health, to the replication of key research studies, especially controlled trials. Replication is the only scientific strategy we have in addressing the inevitable limitations of individual trials no matter how large and well-designed (Iso-Ahola, 2020; Nosek & Errington, 2020).
- RCTs have an important role to play in evaluating creative arts therapies, and arts for health programmes, but qualitative studies are essential too. It should be recognised, however, that neither participants nor professionals facilitating arts activities can be blind to the activity they are engaged in.
- We should recognise the role of personal choice and active agency in engaging with creative activities rather than regarding the arts as a form of treatment.

- Further attention needs to be given to academic curricula in the training of practitioners and researchers in the field of music therapy, and the wider field of arts and health to ensure that a proper critical perspective is adopted in evaluating published research and reviews. This also encompasses solid training in basic statistics, trial designs with their strengths and limitations, sources of bias in data acquisition, and the reasoning behind guidelines such as CONSORT, PRISMA, and others.[17] The process we illustrate here of starting with a piece of research and examining how it is treated in systematic reviews may well be an excellent exercise for post-graduate students in research methods and appraisal.
- Practitioners and researchers in music therapy, and in the wider field of arts and health, should approach systematic reviews and meta-analysis with an appropriate degree of caution.

## REFERENCES


Bazargan, Y., & Pakdaman, S. (2016). The effectiveness of art therapy in reducing internalizing and externalizing problems of female adolescents. *Archives of Iranian Medicine*, *19*(1), 51-56. http://www.ams.ac.ir/AIM/NEWPUB/16/19/1/0010.pdf

Belski, N., Abdul-Rahman, Y. E., Balasundram, V., & Diep, D. (2021). The effectiveness of music therapy in improving depression and anxiety symptoms among children and adolescents – a systematic review. *Child and Adolescent Mental Health*, 1-9. https://doi.org/10.1111/camh.12526

Benjamin, D. J., Berger, J. O., Johannesson, M., & Nosek, B. A. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bosgraf, L., Spreen, M., Pattiselanno, K., & van Hooren, S. (2020). Art therapy for psychosocial problems in children and adolescents: A systematic narrative review on art therapeutic means and forms of expression, therapist behavior, and supposed mechanisms of change. *Frontiers Psychology*, *11*, 584685. https://doi.org/10.3389/fpsyg.2020.584685

Clift, S., Phillips, K., & Pritchard, S. (2021). The need for robust critique of research on the social and health impacts of the arts. *Cultural Trends*, *30*(5), 442-459. https://www.tandfonline.com/doi/full/10.1080/09548963.2021.1910492

Cohen-Yatziv, L., & Regev, D. (2019). The effectiveness and contribution of art therapy work with children in 2018 – what progress has been made so far? A systematic review. *International Journal of Art Therapy*, *24*(3), 110-112. https://doi.org/10.1080/17454832.2019.1574845

Dowlen, R. (2021). *Research digest: young people's mental health*. Leeds: Centre for Cultural Value. https://www.culturehive.co.uk/CVIresources/research-digest-young-peoples-mental-health/

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*(5), 517. https://doi.org/10.1037/0003-066X.33.5.517.a

Eysenck, H. J. (1984). Meta-analysis: An abuse of research integration. *The Journal of Special Education*, *18*(1), 41-59. https://doi.org/10.1177/002246698401800106

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, *309*, 789-792. https://doi.org/10.1136/bmj.309.6957.789

Eysenck, H. J. (1995). Meta-analysis or best-evidence synthesis? *Journal of Evaluation in Clinical Practice*, *1*(1), 29–36. https://hanseysenck.com/wp-content/uploads/2019/12/1995_eysenck_-_meta-analysis_of_best-evidence_synthesis_journal_of_evaluation_in.pdf

Fancourt, D., & Finn, S. (2019). *What is the evidence on the role of the arts in improving health and wellbeing? A scoping review.* World Health Organization. https://www.euro.who.int/en/publications/abstracts/what-is-the-evidence-on-the-role-of-the-arts-in-improving-health-and-well-being-a-scoping-review-2019

Geipel, J., Koenig, J., Hillecke, T. K., Resch, F., & Kaess, M. (2018). Music-based interventions to reduce internalizing symptoms in children and adolescents: A meta-analysis. *Journal of Affective Disorders*, *225*, 647-656. https://doi.org/10.1016/j.jad.2017.08.035

Glew, S. G., Simonds, L. M., & Williams, E. I. (2021). The effects of group singing on the wellbeing and psychosocial outcomes of children and young people: A systematic integrative review. *Arts & Health*, *13*(3), 240-262. https://doi.org/10.1080/17533015.2020.1802604

Goldbeck, L., & Ellerkamp, T. (2012). A randomized controlled trial of multimodal music therapy for children with anxiety disorders. *Journal of Music Therapy*, *49*(4), 395–413. https://doi.org/10.1093/jmt/49.4.395

Grebosz-Haring, K., Schuchter-Wiegand, A. K., Feneberg, A. C., Skoluda, N., Nater, U. M., Schütz, S., & Thun-Hohenstein, L. (2022). The psychological and biological impact of "In-Person" vs. "Virtual" choir Singing in children and adolescents: A pilot study before and after the acute phase of the COVID-19 outbreak in Austria. *Frontiers in Psychology*, *12*, 773227. https://doi.org/10.3389/fpsyg.2021.773227

Grebosz-Haring, K., & Thun-Hohenstein, L. (2018). Effects of group singing versus group music listening on hospitalized children and adolescents with mental disorders: A pilot study. *Heliyon, 4,* e01014. https://doi.org/10.1016/j.heliyon.2018. e01014


[17] See the EQUATOR site for useful guidance: https://www.equator-network.org/

Grebosz-Haring, K., & Thun-Hohenstein, L. (2020). Singing for health and wellbeing in children and adolescents with mental disorders. In R. Hayden, D. Fancourt & A. J. Cohen (Eds.), *The Routledge companion to interdisciplinary studies in singing: Volume III – Wellbeing* (pp.61-73). Routledge.

Grebosz-Haring, K., Thun-Hohenstein, L., Clift, S., Schuchter-Wiegand, A. K., Irons, Y., & Bathke, A. (2021). Effects of arts-based interventions on children and adolescents with mental disorders: a systematic review and meta-analysis. PROSPERO CRD42021193283. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021193283

Grebosz-Haring, K., Thun-Hohenstein, L., Schuchter-Wiegand, A. K., Irons, Y., Bathke, A., & Clift, S. (2022). The need for robust critique of arts and health research: Young people, art therapy and mental health. *Frontiers in Psychology*, *13*, 821093. https://doi.org/10.3389/fpsyg.2022.821093

Greenhalgh, T., Thorne, S., & Malter, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation, 48*, e12931. https://doi.org/10.1111/eci.12931

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., Sterne, J. A. C., Cochrane Bias Methods Group & Cochrane Statistical Methods Group (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

Ioannidis, J.P.A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, *94*(3), 485-514. https://www.milbank.org/wp-content/uploads/2016/10/Milbank_Quarterly_Vol94_Issue3_The_Mass_Production_of_Redundant_Misleading_and_Conflicted_Systematic_Reviews_and_Meta-Analyses.pdf

Iso-Ahola, S. E. (2020). Replication and the establishment of scientific truth. *Frontiers in Psychology*, *11*, 2183. https://doi.org/10.3389/fpsyg.2020.02183

Keller, N. E., Hennings, A. C., & Dunsmoor, J. E. (2020). Behavioral and neural processes in counterconditioning: Past and future directions. *Behaviour Research and Therapy*, *125*, 203532. https://doi.org/10.1016/j.brat.2019.103532

Lu, G., Jia, R., Liang, D., Yu, J., Wu, Z., & Chen, C. (2021). The effects of music therapy on anxiety: a meta-analysis of randomized controlled trials. *Psychiatry Research*, *304*, 114137, 1–13. https://www.sciencedirect.com/science/article/pii/S0165178121004339

MacLure, M. (2005). 'Clarity bordering on stupidity': where's the quality in systematic review? *Journal of Educational Policy, 20*(4), 393–416. https://doi.org/10.1080/02680930500131801

Mansfield, L., Kay, T., Meads, C., Grigsby-Duffy, L., Lane, J., John, A., Daykin, N., Dolan, P., Testoni, S., Julier, G., Payne, A., Tomlinson, A. & Victor, C. (2018). Sport and dance interventions for healthy young people (15–24 years) to promote subjective well-being: A systematic review. *BMJ Open*, *8*, e020959. https://doi.org/10.1136/bmjopen-2017-020959

Møller, M. H., Ioannidis, J. P. A., & Darmon, M. (2018). Are systematic reviews and meta-analyses still useful research? We are not sure. *Intensive Care Medicine*, *44*, 518–520. https://doi.org/10.1007/s00134-017-5039-y

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., & Chou, R. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, *372*, n71 http://dx.doi.org/10.1136/bmj.n71

Page, M. J., Shamseer, L., & Tricco, A. C. (2018) Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Systematic Reviews*, *7*, 32. https://doi.org/10.1186/s13643-018-0699-4

Ponomarenko, A., Yap, J., & Peeran, U. (2017). *Investigating the efficacy of art and music therapy with vulnerable children and young people: A systematic review*. London: Thomas Coram Foundation. https://pearsfoundation.org.uk/wp-content/uploads/2018/06/Coram-Creative-Therapies-Literature-Review.pdf

Ranganathan, P., Pramesh, C.S., & Aggarwal, R. (2016). Common pitfalls in statistical analysis: Intention-to-treat versus per protocol analysis. *Perspectives in Clinical Research*, *7*, 144–6. https://doi.org/10.4103/2229-3485.184823

Robb, S. L., Hanson-Abromeit, D., Maya, L., Hernandez-Ruiz, E., Allison, M., Beloat, A., Daugherty, S., Kurtz, R., Ott, A., Oydele, O. O., Polaski, S., Rager, A., Rifkin, J., & Wolf, E. (2018). Reporting quality of music intervention research in healthcare: A systematic review. *Complementary Therapies in Medicine*, *38*, 24-41. https://doi.org/10.1016/j.ctim.2018.02.008

Schulz, K. F., Altman, D. G., Moher, D. for the CONSORT group (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, *340*, c332. https://doi.org/10.1136/bmj.c332

Seligman, L. D., Ollendick, T. H., Langley, A. K., & Bechtoldt Baldacci, H. (2004). The utility of measures of child and adolescent anxiety: A meta-analytic review of the revised Children's Manifest Anxiety Scale, the State–Trait Anxiety Inventory for Children, and the Child Behavior Checklist. *Journal of Clinical Child and Adolescent Psychology*, *33*(3), 557-565. https://doi.org/10.1207/s15374424jccp3303_13

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A. & the PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *British Medical Journal*, *349*: g7647. https://www.bmj.com/content/bmj/349/bmj.g7647.full.pdf

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review*, *17*(8), 881-901. https://doi.org/10.1016/S0272-7358(97)00056-1

Sharpe, D. & Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology/Psychologie canadienne*, *61*(4), 377–387. https://doi.org/10.1037/cap0000215

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E. & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal*, *358*, 4008. https://doi.org/10.1136/bmj.j4008

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003

Sterne, J.A., Savović, J., Page, M.J., Elbers, R.G., Blencowe, N.S., Boutron, I., Cates, C.J., Cheng, H.Y., Corbett, M.S., Eldridge, S.M., & Emberson, J.R. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *British Medical Journal*, *366*, 4898. https://doi.org/10.1136/bmj.l4898

Strand, N., Fang, L., & Carlson, J.M. (2021). Sex difference in anxiety: An investigation of the moderating role of sex in performance monitoring and attentional bias to threat in high trait anxious individuals. *Frontiers in Human Neuroscience*, *15*, 627589. https://doi.org/10.3389/fnhum.2021.627589

Tripepi, G., Chesnaye, N. C., Dekker, F. W., Zoccali, C., & Jager, K. J. (2020). Intention to treat and per protocol analysis in clinical trials. *Nephrology*, *25*, 513–517. https://doi.org/10.1111/nep.13709

Ελληνική περίληψη | Greek abstract

# Η ανάγκη σθεναρής κριτικής για την έρευνα στις τέχνες και την υγεία: Μία εξέταση της τυχαιοποιημένης ελεγχόμενης μελέτης των Goldbeck και Ellerkamp (2012) για τη μουσικοθεραπεία για το άγχος σε παιδιά, και της αντιμετώπισής της σε τέσσερις συστηματικές ανασκοπήσεις

Stephen Clift | Katarzyna Grebosz-Haring | Leonhard Thun-Hohenstein | Anna Katharina Schuchter-Wiegand | Arne C. Bathke

## ΠΕΡΙΛΗΨΗ

Περιγράφουμε μία εργασία σε εξέλιξη για τη διεξαγωγή μίας συστηματικής ανασκόπησης ερευνών που αφορούν τον αντίκτυπο προγραμμάτων βασισμένων στις τέχνες στην ψυχική υγεία νέων ανθρώπων. Αναζητήσαμε σχετικές μελέτες μέσω κύριων βάσεων δεδομένων και εξετάσαμε υπάρχουσες συστηματικές ανασκοπήσεις για επιπλέον μελέτες οι οποίες πληρούν τα κριτήρια ένταξης στην έρευνά μας. Έχουμε ωστόσο επιφυλάξεις τόσο ως προς την ποιότητα των υφιστάμενων αρχικών μελετών, όσο και ως προς τις πρόσφατα δημοσιευμένες συστηματικές ανασκοπήσεις σε αυτό το πεδίο των τεχνών και της υγείας. Σε προηγούμενο άρθρο (Grebosz-Haring et al., 2022) εστιάσαμε σε μία τυχαιοποιημένη ελεγχόμενη δοκιμή (ΤΕΔ) για την εικαστική θεραπεία για έφηβες με «εσωτερικευμένα» και «εξωτερικευμένα» προβλήματα, και την συμπερίληψη αυτής της δοκιμής σε τρεις συστηματοποιημένες ανασκοπήσεις, και εκφράσαμε τους προβληματισμούς μας. Σε αυτό το άρθρο, επεκτείνουμε το πεδίο της κριτικής μας εξέτασης σε μία μελέτη που αφορά στη μουσικοθεραπεία με παιδιά που αναφέρεται ότι αντιμετωπίζουν αγχώδεις διαταραχές (Goldbeck & Ellerkamp, 2012), και το πώς χρησιμοποιήθηκε αυτή η μελέτη σε τέσσερις πρόσφατες συστηματικές ανασκοπήσεις / μετα-αναλύσεις (Ponomarenko et al., 2017; Geipel et al., 2018; Lu et al., 2021; Belski et al., 2021). Παρουσιάζουμε τους περιορισμούς της μελέτης των Goldbeck και Ellerkamp που υποσκελίζουν το συμπέρασμα στο οποίο καταλήγουν για την αποτελεσματικότητα της μουσικοθεραπείας στην ύφεση των αγχωδών διαταραχών. Επίσης καταδεικνύουμε ότι οι ανασκοπήσεις δεν είναι επαρκώς κριτικές και αντιμετωπίζουν με λανθασμένο τρόπο την έρευνα των Goldbeck και Ellerkamp, κάτι που δημιουργεί αμφιβολίες ως προς την αξιοπιστία τους. Καταληκτικά, συλλογιζόμαστε ως προς τα μαθήματα που αποκομίσαμε από τη δική μας κριτική και χαράζουμε κάποιες θετικές προτάσεις για μελλοντικές έρευνες και την διεξαγωγή ανασκοπήσεων.

## ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

μουσικοθεραπεία, παιδιά, άγχος, συστηματική ανασκόπηση, μετα-ανάλυση, κριτική