DOI: 10.1111/nyas.15006

REVIEW

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

The need for robust critique of arts and health research: Dance-movement therapy, girls, and depression

Katarzyna Grebosz-Haring[1,2] | Leonhard Thun-Hohenstein[3] | Anna K. Schuchter-Wiegand[1,2] | Arne C. Bathke[4] | Stephen Clift[5]

[1]Interuniversity Organisation Science & Art, Paris Lodron University Salzburg/Mozarteum University Salzburg, Salzburg, Austria

[2]Faculty of Art History, Musicology and Dance Studies, Paris Lodron University Salzburg, Salzburg, Austria

[3]Paracelsus Medical University, Salzburg, Austria

[4]Faculty of Artificial Intelligence and Human Interfaces, Paris Lodron University Salzburg, Salzburg, Austria

[5]Sidney De Haan Research Centre for Arts and Health, Canterbury Christ Church University, Canterbury, UK

Correspondence
Katarzyna Grebosz-Haring, Interuniversity Organisation Science & Art, Faculty of Art History, Musicology and Dance Studies, Paris Lodron University Salzburg, Mozarteum University Salzburg, Bergstrasse 12, 5020 Salzburg, Austria.
Email: katarzyna.grebosz-haring@plus.ac.at

Funding information
Salzburg Land

Abstract

We examine a highly cited randomized controlled trial on dance-movement therapy with adolescent girls with mild depression and examine its treatment in 14 evidence reviews and meta-analyses of dance research. We demonstrate substantial limitations in the trial which seriously undermine the conclusions reached regarding the effectiveness of dance movement therapy in reducing depression. We also show that the dance research reviews vary substantially in their treatment of the study. Some reviews provide a positive assessment of the study and take its findings at face value without critical commentary. Others are critical of the study, identifying significant limitations, but showing marked differences in Cochrane Risk of Bias assessments. Drawing on recent criticisms of systematic reviewing and meta-analysis, we consider how reviews can be so variable and discuss what is needed to improve the quality of primary studies, systematic reviews, and meta-analyses in the field of creative arts and health.

KEYWORDS
adolescent girls, arts and health, dance-movement therapy, depression, meta-analyses, reviews

INTRODUCTION

The field of arts therapy and arts and health research has developed considerably since the beginning of the century. However, Clift et al.[1] evaluate two recent scoping reviews on the arts and health literature that were commissioned by the World Health Organization (WHO) and the UK Department for Digital, Culture, Media, and Sport (DCMS).[2,3] They document problems associated with a lack of critical evaluation on the research that was included in those reviews. The positive conclusions and recommendations drawn from these reviews are called into question, and Clift et al.[1] suggest that

> …it is premature to suggest, as the WHO and DCMS reports do, that the evidence on arts and health provides a secure foundation on which to develop social and health policy. In moving research and practice forward in future, the field must rely on rigorous systematic reviews involving careful quality assessment of both quantitative and qualitative studies. (p. 13)

Guided by this view, we planned to undertake a systematic review of controlled studies of creative arts activities/arts therapy for children and young people experiencing mental health problems based

on earlier pilot studies in Salzburg.[4–6] However, our preparatory work revealed some concerns. The first regarded the quality of the published research on the effects of arts programs and therapy for young people with mental health challenges, and the second was a lack of critical evaluation in recent reviews of this literature.

Quality of primary trials, systematic reviews, and meta-analyses

In addition to placing our work in the context of research on creative arts and health, we also draw on methodological guidelines and critical discussions of medicine and healthcare research and evidence reviews. The last 40 years have seen considerable growth of guidance for the conduct and reporting of controlled trials (CONSORT[7–9]), assessments of risks of bias in trials,[10] the development of the Cochrane Risk of Bias (RoB) tool,[11–13] and the reporting and evaluation of systematic reviews and meta-analyses (PRISMA,[14,15] AMSTAR-2[16,17]).

Systematic reviews and meta-analyses are widely considered as being at the top of most models of evidence hierarchies[18] and meta-analyses have even been characterized as providing the platinum standard in the synthesizing of evidence.[19] However, reservations have been expressed about the principles and practice of systematic reviews and meta-analysis and their weaknesses in their execution. MacLure,[20] for example, presents a critique of the systematic reviews on educational topics that were conducted and supported by the EPPI-Centre at the University of London over the period of 2002–2004. In health research, Greenhalgh et al.[21] are critical of the view that systematic reviews are necessarily superior to narrative reviews. Ioannidis[22] argues that most systematic reviews and meta-analyses are "unnecessary, misleading, and/or conflicted" (p. 468). Møller et al.[23] argue that most systematic reviews and meta-analyses are "focused on unimportant questions…redundant and unnecessary…flawed beyond repair…only about 3% of them [were] well done and clinically useful" (p. 520).

Moreover, Eysenck[24,25] argues that the data summarized in meta-analyses should be homogeneous—that is, patients, treatments, and outcomes must be similar or at least comparable, and studies should be methodologically sound. Yet, often there is evidence of heterogeneity whereby reviewers are adding apples to oranges and including studies of variable methodological quality. Reservations have continued ever since despite attempts to tackle these early criticisms.[26,27] Stegenga[19] 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). These concerns will be explored further in our discussion.

Previous critiques of art therapy randomized control trials and their treatment in evidence reviews

In a previous paper,[28] we found substantial limitations in a randomized control trial (RCT) of art therapy for adolescent girls with internalizing or externalizing problems[29] and a lack of critical evaluation in three systematic reviews which included it. In a second paper,[30] we identified significant problems with an RCT of music therapy for children with anxiety disorders[31] and found that two systematic reviews and two meta-analyses took the findings from this trial at face value with little critical evaluation.

In this paper, we repeat and extend this analysis, starting with an RCT reported by Jeong et al.,[32] on the effects of dance movement therapy (DMT) on depression in adolescent girls in South Korea and examine its treatment in subsequent evidence reviews.

Our objectives are:

- To provide a nonevaluative account of the methods, findings, and limitations of the Jeong et al. trial[32] as presented by the authors.
- To offer a robust critique of the Jeong et al. trial[32] and identify substantial concerns.
- To examine the treatment of the Jeong et al. research[32] in 14 evidence reviews which focus on evaluations of DMT and dance research.
- To discuss our findings in light of recent criticisms of systematic reviewing and meta-analysis and offer recommendations for future research and reviewing.

METHODS

A protocol[33] was developed for a systematic review of controlled studies on creative arts interventions for children and young people experiencing challenges to their mental health. Seven databases were systematically searched, and two members of the team (K.G.-H. and S.C.) independently screened abstracts for relevance. Full-text papers were obtained and organized alphabetically by A.K.S.-W. and the first author for a further assessment of study relevance for the systematic review. In addition, the citation function of Google Scholar was used to identify relevant subsequent publications referring to the papers identified from our primary search. Google Scholar searches served a valuable function in identifying further evidence reviews of relevance to our focus on arts interventions for children and young people. This paper focuses on a specific RCT on DMT to address mild depression in adolescent girls[32] and how that RCT is treated in subsequent evidence reviews identified through the use of Google Scholar. This RCT was identified through our search strategy for the systematic review referred to above.[33]

A nonevaluative summary of the Jeong et al. RCT[32] will be followed by a careful analysis of the design and methods of the study, the procedures for assessing the participants, the description of the intervention and the quality of the analysis and reporting of findings, and the credibility of conclusions drawn. This will be guided by standards for the conduct and reporting of trials that were in place prior to the date of the study,[10] as well as the continued development of these standards.[7–9,11–13] Fourteen dance-focused evidence reviews will be evaluated with reference to established standards for the design and conduct of systematic reviews (PRISMA and preregistration[14]), and the AMSTAR-2 checklist will be used to asses review quality.[17]

RESULTS

A Google Scholar search (November 24, 2022) showed that the Jeong et al. RCT[32] had been cited 386 times in further publications and was included in 25 evidence reviews published in peer-reviewed journals between 2011 and 2022. These reviews vary in the character of interventions they consider: 14 reviews focus on dance activities, including DMT; seven concern exercise/physical activity and so consider the DMT intervention in the Jeong et al. trial[32] as a form of exercise; and four address a wide range of psychological treatments and therapies, including creative arts therapies. In this paper, we limit our attention to the treatment of the Jeong et al.[32] research in dance-focused research reviews. This is justified given our interest in a broad range of creative arts interventions, including arts therapies for children and young people.

Jeong et al.: A nonevaluative summary

Jeong et al.[32] identified 40 girls (mean age was 16 years) in a single school in South Korea, with "mild depression" based on the Beck Depression Inventory (BDI)[34] and further assessments. The girls were randomly allocated to 12 weeks of DMT ($n = 20$) or a waiting control ($n = 20$). Sessions of 45 min took place three times per week, so the girls experienced 36 sessions and a total of 27 h of DMT. As a measure of psychological distress, Jeong et al.[32] used the Symptom Checklist-90-Revision (SCL-90-R)[35] at baseline and after 12 weeks. This consists of 90 questions giving rise to nine subscales (one of which is depression) and three summary indices. In addition, plasma concentrations of cortisol, serotonin, and dopamine were assayed before and after the study. Repeated measures ANOVA revealed significant time × group interactions for all subscale and global scores on the SCL-90-R, indicating improved scores in the DMT group compared with the control. Plasma serotonin concentration increased, and dopamine concentration decreased in the DMT group, but no changes were found for cortisol. The authors conclude that their data "suggest that DMT has relaxation effects, stabilizes the sympathetic nervous system, and may be beneficial in improving the symptoms of mild depression." They acknowledge, however, that their study was preliminary, had a small sample size, and lacked an equivalent exercise control group to estimate a possible expectation effect.

Jeong et al.: A robust critique

There are serious limitations with the Jeong et al.[32] research and paper, few of which are acknowledged in the reviews we consider below. We have concerns about the lack of a clear rationale for the study based on an appropriate theoretical perspective and evidence review. For example, the authors offer no review of existing evidence that regular dance (or physical activity or exercise) can lead to a reduction in depression (see Refs. [36] and [37]) or modulate neurotransmitter and endocrine levels. Also, no reference is made to CONSORT guidelines for reporting RCTs, nor is a CONSORT checklist or flow diagram provided despite the publication of the first CONSORT statement in 1996[38]—9 years before this study. More importantly, the study has five serious methodological limitations, which we identify in detail.

Uncertainty regarding the "depressed" status of the girls taking part

It is not clear how the authors determined that the girls were "mildly" depressed, nor whether they were all in the "mild" category or whether they were "at least" mildly depressed, with some experiencing more serious issues. A sample of 300 girls in a middle school in Iksan, South Korea completed the BDI, and 112 girls "with higher depression scores were selected as possible subjects." However, the scores on the BDI for this sample of 112 girls are not reported and no cutoff points for mild depression are given. A further selection of 75 girls was then followed based on six criteria, including: "no past or present diagnosis of psychiatric or internal illnesses" and "not using prescription medication or any other therapeutic treatment for depression." We wonder, therefore, in what sense the 75 girls selected at this stage were "depressed"? However, in a further step of sample selection, "potential subjects underwent a pre-treatment assessment of symptoms over four weeks to confirm a *diagnosis* [our emphasis] of depression" (p. 1714). The nature of this further assessment is not described, but it resulted in the exclusion of 24 girls "because they could not be diagnosed as having symptoms of depression." Of the 51 girls regarded as depressed at this stage, 40 were randomly selected and then randomly assigned to either DMT or a waiting list control. In the assessment of the outcomes of the trial, no data from the BDI are reported, but instead, use is made of the SCL-90-R scales.

Inadequate account of the randomization process

Jeong et al.[32] state that the girls were randomly assigned to the treatment and control groups "by a secretary who was blind to the experimental procedure" and no details are provided of the method employed. In terms of the criteria specified in the guide for the Cochrane RoB tool,[12,13] the description of the process should indicate, at a minimum, that randomization was done and in such a way that it was concealed from the research team. Group differences in SCL-90-R scores apparent at baseline were not tested for statistical significance by Jeong et al.[32] but may be compatible with randomization. A benefit of the doubt judgment would be that the trial had a low risk of bias due to problems with randomization, or to be more cautious, that this is unclear. The operation of bias, however, would almost certainly be in the direction of enhancing the treatment effect.

Problems with the control condition

The girls in this study were attending a single school, and consequently, those assigned to the control group would have been aware that girls in the DMT group were participating in dance activities three times a week over 12 weeks. The potential for "demoralized resentment"

among the control group cannot be discounted. Cunningham et al.[39] suggest, for example, that "waiting list control designs in psychological and behavioral intervention research may artificially inflate intervention effect estimates." Weisz et al.[40] in a major review of five decades of research on youth psychotherapies also point out that waitlist controls tend to inflate the treatment effect of psychotherapies and that "usual care may be a particularly rigorous standard of comparison" (p. 94).

Jeong et al.[32] acknowledge that a limitation of their study is the lack of an active control condition to take account of a possible expectation effect. In their discussion, they suggest that "an equivalent exercise control group" would have been appropriate, but they offer no explanation for why they did not employ one. If the girls had been given a form of psychiatric assessment which indicated that treatment was needed, then all participants in the trial should have been offered standard medical treatment, with DMT as an adjunct therapy. The control group, in other words, would have received treatment as usual, which is the generally recommended approach for medical trials of a new intervention.[40]

Limited details of the administration, scoring, and interpretation of the SCL-90-R

The authors state that all participants "completed a self-report inventory of emotional distress, the Symptom Check List-90-Revision (SCL-90-R)" which has nine subscales and three summary indices. However, it is not clear whether this was part of the preassessment referred to above or if participants completed the questionnaire once they were identified as showing symptoms of depression. It is also not explicitly stated whether the SCL-90-R was completed prior to randomization. Jeong et al.[32] go on to say that "the raw scores on the SCL-90-R have been converted to standard T scores and normalized to the non-patient population of Korea" (p. 1714). However, no source is provided for the Korean norms.

The manual of the SCL-90-R, and other sources make clear, however, that raw scores for subscales can be transformed into T-scores for clinical use[41,42] with a mean of 50 and standard deviation of 10. Results for the nine subscales and three summary indices derived from the SCL-90-R are reported by Jeong et al.[32] in their Table 2 (p. 1716). All mean values reported for the DMT and control group at baseline are close to the population mean of 50 (range 43–57). The depression mean score for the DMT group is 51.8 and 43.6 for the control group. These mean scores are, respectively, just above and clearly below the standardized population mean, and neither suggest that the girls were depressed. In addition, it is striking that mean values for all the subscales are similar, suggesting that if the girls were "depressed" at baseline, they were also affected by somatic symptoms and were anxious, obsessive-compulsive, interpersonally sensitive, hostile, phobic, paranoid, and psychotic. The mean global severity index, based on a sum of responses to all items in the questionnaire, is reported as 51.3 for the DMT group and 44.5 for the control group at baseline. Again, in the absence of a clinical cutoff value, there is no evidence that the girls in the study were sufficiently distressed to warrant a therapeutic intervention.

Lack of detail regarding the assessment of cortisol, serotonin, and dopamine

Jeong et al.[32] give no information on how, and more importantly, when, blood samples were taken to assess levels of cortisol, serotonin, and dopamine. Timing is especially crucial with respect to cortisol levels as this hormone has a marked diurnal cycle.[43] Levels of neurotransmitters and cortisol are also likely to be sensitive to factors, such as exercise, diet, consumption of coffee, and experience of stress, prior to assessment.[44] This lack of detail is particularly damaging to the scientific credibility of the Jeong et al. trial,[32] as it renders their work nonreplicable. In addition, Jeong et al.[32] do not state explicitly that the researchers gathering and analyzing serum samples for neurohormone concentrations were blind to treatment condition.

The results for the assaying of serotonin, dopamine, and cortisol are not presented in table form, but rather as three graphs (fig. 1, p. 1718). From these, it is possible to discern the quantification of concentrations of each molecule and their mean values, but the vertical lines above and below the means are large, overlapping, and are not explained—they may represent confidence intervals, standard deviations, or standard errors. Significant group × time interactions are claimed for serotonin and dopamine, but no such interaction is said to emerge for cortisol even though the graph shows that cortisol increased in the control group and slightly declined in the DMT group. In the absence of directly reported data, it is difficult to judge the validity of the analysis reported and thus the conclusions drawn. In addition, no attempt is made to explain why serotonin increased, but dopamine decreased, and why cortisol remained unchanged. There is no explanation offered as to how these changes relate to the (suggested) activity of the sympathetic nervous system or to changes in SCL-90-R scores. Jeong et al.[32] presumably mean that dance leads to a shift in autonomic tone—the balance between the actions of the sympathetic and parasympathetic autonomic nervous systems[45]—but they do not provide further details.

Treatment of Jeong et al. in subsequent dance-focused reviews

As noted earlier, the Jeong et al.[32] trial is included in 25 evidence reviews published between 2011 and 2022. Of relevance to this paper are reviews which focus on dance and DMT programs, and excluded from consideration are reviews that focus on physical activity and exercise (see, e.g., Pascoe et al.[36,37]). Evidence reviews come in a variety of forms,[46] and we have not limited the character of dance reviews considered, taking only their reference to the Jeong et al. paper[32] as a basis for inclusion in this paper.

An overall assessment of review quality was undertaken (see Table 1)2 based on whether the review was registered in PROSPERO or COCHRANE, whether PRISMA guidelines were followed in the conduct of the review, and whether AMSTAR-2 criteria were met. In Table 1, Green = yes, Yellow = partially met, and Red = no.[20] Reviews are ordered from left to right by quality. Four reviews present

TABLE 1 Fourteen dance-focused evidence reviews that include Jeong et al.[32]—Preregistration, PRISMA, and AMSTAR-2 ratings.

Author	Koch et al.[47]	Koch et al.[48]	Meek-ums et al.[49]	Karkou et al.[50]	Strassel et al.[51]	Kiepe et al.[52]	Millman et al.[53]	Tao et al.[54]	Mala et al.[55]	Calçada and Gilham[56]	Grudzi-ńska and Izdebski[57]	Lopez-Nieves and Jako-bsche[58]	Lossing et al.[59]	Lykesas et al.[60]
Character of review	Meta-analysis	Meta-analysis	Coch-rane review	Meta--analyses	Syste-matic review	Syste-matic review	Syste-matic review	Syste-matic review	Scoping review	Scoping review	Evid-ence review	Evid-ence review	Evid-ence review	Evid-ence review
PROSPERO (P), Cochrane (C), or an update (Up)	No	Up	C	Up	No	No	No	P	N/A	N/A	N/A	N/A	N/A	N/A
PRISMA guidelines referenced	No	No	No	Yes	No	No	Yes	Yes	N/A	Yes	N/A	N/A	N/A	N/A
PRISMA style diagram	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No
1. Research questions and inclusion criteria include PICO														
2. Statement that methods were established prior to review*														
3. Explains selection of study designs for inclusion														
4. Comprehen-sive literature search strategy*														
5. Study selection in duplicate														
6. Data extraction in duplicate														
7. Lists excluded studies and justifies exclusions*														
8. Describes included studies in adequate detail														
9. Assesses risk of bias (RoB)*														
10. Sources of funding for the studies included														
11. Meta-analysis uses appropriate methods*														

(Continues)

TABLE 1 (Continued)

Author	Koch et al.[47]	Koch et al.[48]	Meek-ums et al.[49]	Karkou et al.[50]	Strassel et al.[51]	Kiepe et al.[52]	Millman et al.[53]	Tao et al.[54]	Mala et al.[55]	Calçada and Gilham[56]	Grudzi-ńska and Izdebski[57]	Lopez-Nieves and Jako-bsche[58]	Lossing et al.[59]	Lykesas et al.[60]
12. Meta-analysis assesses the impact of RoB in studies														
13. Accounts for RoB when interpreting the results*														
14. Provides a satisfactory account of heterogene-ity														
15. Quantitative synthesis investigates publication bias*														
16. Conflicts of interest, including any funding														
Critical flaws and quality assessment	0 High	0 High	0 High	1 High	2 Mid	2 Mid	3 Low	1 High	3 Low	3 Low	3 Low	4 Low	4 Low	3 Low

Notes: * = Items in AMSTAR-2 that can "critically affect the validity of a review and its conclusions."[17] Green (gray in print) = yes, Yellow (light gray in print) = partial yes, Red (dark gray in print) = no, White = not applicable.
Abbreviations: AMSTAR-2, A MeaSurement Tool to Assess systematic Reviews 2; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROSPERO, International Prospective Register of Systematic Reviews; RoB, Risk of Bias.

TABLE 2 Cochrane RoB assessments of Jeong et al.[32] in three meta-analyses and one systematic review.

RoB criteria (Higgins et al.[11])	Koch et al.[47]	Meekums et al.[49]	Karkou et al.[50]	Tao et al.[54]
Random sequence generation (selection bias)				
Allocation concealment (selection bias)				
Blinding of participants and personnel (performance bias)				
Blinding of outcome assessment (detection bias)				
Incomplete outcome data (attrition bias)				
Selective reporting (reporting bias)				
Other bias				
Overall assessment of bias (and direction)				

Note: Green (gray in print) = low risk, Yellow (light gray in print) = unclear risk, Red (dark gray in print) = high risk, White = no rating.
Abbreviation: RoB, Risk of Bias.

a meta-analysis,[47–50] four reviews are systematic,[51–54] two are scoping reviews,[55,56] and the remaining four are described by the authors as reviews.[57–60]

One review protocol was registered in Cochrane,[49,61] and only one protocol in PROSPERO.[54] Two reviews served to update previous reviews.[48,50] Four systematic reviews were not preregistered in PROSPERO.[47,51–53] Two scoping reviews[55,56] and a further four evidence reviews[57–60] were not eligible for registration in PROSPERO. One meta-analysis,[50] two systematic reviews,[53,54] and one scoping review[56] referred to PRISMA guidelines, but four further reviews[48,49,51,52] did include a PRISMA-style flow diagram of progress in the review.

AMSTAR-2[17] is a screening tool for assessing the quality of systematic reviews and meta-analyses. Detailed guidance is provided for judging reviews against 11 criteria for systematic reviews and a further five criteria for meta-analyses. Seven of these criteria are regarded by the developers of AMSTAR-2 as "critical" to the quality of a review. AMSTAR-2 was applied to all the reviews in the interests of consistency and to highlight the strengths and limitations of each review. Assessments were made by SC and independently moderated by KGH with any disagreements resolved in discussion. Overall quality was judged by the number of critical criteria that each of the reviews failed to meet: high (0–1 not met), mid (2 not met), and low (3+ not met). The four meta-analysis papers and one systematic review were judged to be high in quality, two systematic reviews received a mid-quality rating, and the remaining seven reviews were rated as low in quality.

The seven low-quality reviews do not meet the standards of systematic reviews and will be considered briefly. Grudzińska and Izdebski[57] are inaccurate in their description of the Jeong et al. study[32] and offer no critical comments. Mala et al.[55] describe Jeong et al.[32] as one of the two "strongest studies of D/MT for clients with depression" in their review but take the Jeong et al.[32] findings at face value. Calçada and Gilham,[56] Lossing et al.,[59] and Lykesas et al.[60] provide a descriptive account of Jeong et al.[32] but no critical analysis. Millman et al.[53] take the Jeong et al.[32] findings at face value but do express skepticism over the idea that changes in neurotransmitter concentrations explain the claimed effects of DMT on depression as "they did not report correlations between these changes" (p. 30). Lopez-Nieves and Jakobsche[58] accurately summarize the Jeong et al. study[32] and go beyond Millman et al.[53] in pointing out that direct correlations cannot be drawn between blood serotonin and dopamine levels and effects in the brain because neither serotonin nor dopamine synthesized in the body can cross the blood–brain barrier.

Two mid-quality evidence reviews were conducted systematically and are an improvement on the seven low-quality reviews, but they contain errors in their treatment of the Jeong et al.[32] trial. Kiepe et al.[52] refer to "a decrease of symptoms and disease severity," whereas Jeong et al.[32] describe the girls in their study as experiencing "mild depression." However, Kiepe et al.[52] identify two strengths—the use of validated questionnaires (the BDI and the SCL-90-R scales) and objective biological markers (although they offer no commentary on the meaning and limitations of these measures). They also identify two weaknesses—a limited description of the DMT program and a lack of formal testing of differences between the dance and control group at baseline. Kiepe et al.[52] are correct in their comment on baseline differences, but Jeong et al.[32] report that a repeated measures ANOVA showed no main effect for group for the three global score profiles from the SCL-90-R.

Strassel et al.[51] consider both earlier reviews on dance therapy and RCTs (including Jeong et al.[32]). Trials were assessed using the Jadad scale,[62] which rates the quality of RCTs on a scale of 0–5, with scores above three indicating a well-designed trial. Criteria include whether the study was randomized, double-blinded, and described dropouts. Strassel et al.[51] acknowledge that the Jadad scale[61] is of limited value for studies of creative arts therapies as blinding is not possible. None of the 18 trials included in the review achieved a score above three, and Jeong et al.[32] was rated at two with randomization described but not appropriately conducted, and dropouts described but insufficient detail given (see tab. 9, p. 58). The reference to dropouts is inaccurate, however, as there was no reported attrition. Nevertheless, the study is said to show that "negative psychological symptoms improved significantly in experimental group but not in the control group" (p. 56). There is no mention of the physiological assessments and no further critical commentary in the text.

We now turn to the five systematic reviews/meta-analyses which were competently conducted according to AMSTAR-2 criteria. Tao et al.[54] in the most recent systematic review give a descriptive account of the methods and findings of the Jeong et al. study.[32] Critical scrutiny of all the included research studies is undertaken using the Cochrane RoB Tool.[19] Most risk ratings for Jeong et al.[32] are "unclear" or a "high risk" rating is given for nonblinding of participants and personnel, which is inevitable in creative arts intervention trials. Surprisingly, Tao et al.[54] make no reference to the earlier meta-analyses reported by Koch et al.,[47,48] Meekums et al.,[49] or Karkou et al.[50]

Koch et al.[47] provide a detailed review of research evidence from 23 primary trials on the psychological outcomes of DMT and dance activities for diverse participants. No formal quality screening was used, but an assessment was made of the included studies in terms of specified aims and objectives, descriptions of participants, specification of inclusion and exclusion criteria, randomization, description of the intervention, reporting of baseline data, and the nonblinding of participants. Overall, Koch et al.[47] are content that "all of the included studies offered a quite satisfactory degree of methodological quality" but they accept that studies did vary in quality "especially with regard to randomization, blinding strategy, and the analysis of baseline differences" (p. 57). The latter issue is relevant to the Jeong et al. study[32] as Koch et al.[47] exclude Jeong et al.[32] from the reported meta-analyses for clinical outcomes, depression, and anxiety because of apparent pretest differences on these variables. Nevertheless, the Jeong et al. findings[32] are included in analyses of change scores, which support the view that DMT positively affects psychological outcomes. Little criticism is offered of the Jeong et al.[32] research in the Koch et al. review[47] apart from identifying a lack of attention to pretest differences on some of the variables they assess. Such differences are not necessarily a source of bias in RCTs if baseline values are used as a covariate or if main effects and group × time interaction effects are carefully considered. As previously noted, ANOVA showed no significant main effects for group for the psychological distress scores between the DMT and waiting control groups.

Koch et al.[48] report an update of an earlier review and meta-analysis.[47] They do not consider the Jeong et al. study[32] in detail as it was included in their earlier review. Nevertheless, in the introduction to this review, they describe the Jeong et al. study[32] as "a high-quality primary trial" and selectively cite the results for dopamine and serotonin without mentioning the findings for psychological symptoms.

In a meticulous Cochrane review, Meekums et al.[49] focus on controlled trials which examined "the effects of DMT for depression with or without standard care" in comparison to a range of control

conditions. Jeong et al.[32] is one of only three studies included in the review. In marked contrast to Koch et al.,[48] Meekums et al.[49] describe the Jeong et al. trial[32] as having "very low methodological quality" (p. 23) and identify multiple problems. They note that it is unclear why the BDI was used initially to identify participants, but then was dropped and results from the SCL-90-R are reported. They also point out that the SCL-90-R scores on the depression subscale were higher at baseline for the DMT group than the control group. Meekums et al.[49] assess risks of bias in the Jeong et al. trial[32] using the Cochrane RoB scale[11] and give high ratings of risk for problems with randomization and allocation concealment, lack of blinding of participants and therapists, and lack of blinding in outcome assessments.

Karkou et al.[50] build upon the Meekums et al.[49] Cochrane review by considering observational DMT studies in addition to RCTs. Karkou et al.[50] accurately report the psychological outcome measures used by Jeong et al.[32] in their Table 2 of study characteristics (p. 11) but offer no commentary on these measures and their limitations. As the focus of the systematic review is depression, no discussion is given of the findings for neurotransmitters and their potential relevance for understanding the psychological impact of dance. As in the Meekum et al.[49] review, the RCTs included were assessed for risk of bias using the Cochrane criteria,[11] and Jeong et al.[32] is again judged to be of "very low methodological quality" (p. 17) due to "high risk of bias" from problems with randomization, allocation concealment and blinding of outcome assessment, and unclear bias linked to incomplete outcome data. The only difference from Meekums et al.[49] is that "blinding of participants and personnel" was omitted from the assessments of bias, as blinding is not possible in studies of therapeutic interventions. However, the unavoidability of blinding does not mean that a study is no longer at risk of bias, especially due to participant expectation and social desirability effects.

Table 2 compares the Cochrane RoB assessments[11] of the Jeong et al.[32] trial in four high-quality reviews. For Koch et al.,[47] the risk of bias assessments given here are judgments we have made based on the discussion in their text. No formal risk of bias assessment was performed and their rationale for not doing so is given. The variations in judgments are striking, indicating little overall consensus.

For each of these reviews, the Cochrane RoB criteria are applied globally to the studies considered. This is incorrect according to current Cochrane standards, as the developers of the current RoB2 tool are explicit that assessments for some criteria need to be made specifically for each outcome measure.[13] In the Jeong et al. study,[32] for example, there is clearly a risk of bias with the SCL-90-R scales both with respect to the lack of blinding of participants and personnel as well as the lack of blinding in outcome assessment. On the other hand, there may be little or no risk of bias for the assaying of neurohormones.

DISCUSSION

This paper has taken a critical look at a target RCT on DMT for adolescent girls with mild depression[32] and examined the assessment of this study in 14 subsequent evidence reviews/systematic reviews/meta-analyses of dance research. This is the third paper following this methodology.[28,30] We acknowledge that some of the reviews considered are not strictly "systematic"—but it is instructive to look at all of them critically, applying current standards for systematic reviews.[16,17] Seven of the reviews[53,55–60] were rated as low in quality and two as mid-quality.[51,52] However, five reviews[47–50,54] are of high quality judged by the AMSTAR-2 criteria.[17] Even so, there are marked contrasts among these careful reviews in their assessments of the Jeong et al. trial.[32] Koch et al.,[47,48] for example, describe the Jeong et al. study[32] as "a high-quality primary trial," whereas Meekums et al.[49] and Karkou et al.[50] describe the methodological quality of the Jeong et al. trial[32] as very low and compromised by high levels of bias. Tao et al.,[54] in contrast, suggest that most risks of bias are unclear. Four high-quality reviews vary considerably in their ratings of the Jeong et al. trial[32] using the Cochrane RoB tool.

Our study adds to the critical literature on systematic reviews and meta-analyses[22,63] and raises four important questions:

Why so many reviews?

In this paper, we have considered 14 reviews which include Jeong et al.[32] and focus on DMT and dance. However, as we noted earlier, 11 of the reviews focused on physical activity and exercise (e.g., Ref. 36) or a wider range of psychological therapies (e.g., Ref. 40), and included Jeong et al.[32] There are some obvious answers as to why there should be so many reviews. The reviews cover a period of 12 years from 2011 to 2022, and new research has appeared to warrant a new review. Furthermore, each review is concerned with a different specific research question so may cover a different body of literature. The clearest justification for repeated reviews is provided by the sequence of papers by a common core of authors, starting with a scoping review,[55] proceeding to a Cochrane protocol,[61] then to a Cochrane Review,[49] and finally to a wider-ranging systematic review with some meta-analysis.[50]

A further factor worth noting is several recent reviews appear to have been conducted without awareness of previous reviews. This is true for Grudzińska and Izdebski,[57] Lopez-Nieves and Jakobsche,[58] Lykesas et al.,[60] and Tao et al.,[54] all of which fail to reference even the major Cochrane Review by Meekums et al.[49] As Siontis and Ioannides[63] note:

> Most systematic reviews are never registered despite the availability of platforms for prospective registration such as PROSPERO. Thus, many teams working on the same topic concurrently may have no knowledge of each other's work-in-progress. However, even when potentially overlapping systematic reviews are published without temporal proximity to each other, authors commonly do not even acknowledge the existence of prior systematic reviews. (p. 2)

Nevertheless, there are more reviews than appear warranted by the size and quality of the corpus of original studies on DMT and

dance.[22,23] It is especially clear that the Jeong et al. trial[32] has been reviewed repeatedly since it appeared and only rarely have substantial limitations of the study been recognized.

Why are reviews so variable in quality?

Not all the reviews are "systematic" as is currently understood and some may be less rigorous as a result. The date of publication does not appear to be a factor, as the earliest reviews[51,52] are satisfactorily conducted. Increasingly, we might expect to see that even scoping or narrative reviews would provide clear details of their search strategy and ensure that selection and data extraction are conducted by at least two reviewers independently. It is surprising to see that some recent reviews[56–58] and even systematic reviews[52,53] do not meet these standards. A further factor may be a limited range of expertise among the team of authors of reviews which, in addition to subject specialists, should include experts in trial design, statistics, and quantitative synthesis.

Why do the meticulous reviews vary in their RoB assessments?

Table 2 demonstrates that even among the most meticulous of systematic reviews and meta-analyses, there is considerable variation in the assessments of risks of bias for the Jeong et al. trial.[32] This is the case even though the authors of these four reviews state that study selection, data extraction, and bias assessments were conducted independently by two reviewers with moderation in the event of disagreement. The variations may indicate one or more of the following:

- The guidance that accompanies the Cochrane RoB Tool employed in the reviews is not sufficiently clear.[13]
- The guidance is clear, but the reviewers have not been thoroughly trained in using the tool.[13]
- Risk of bias assessments are unavoidably subjective.[19]

These reflections are vindicated by the considerations that resulted in the revision of the original Cochrane RoB Tool (used in several of the reviews considered here) to produce a second version (RoB-2). Sterne et al.[13] note:

> After nearly a decade of experience of using the RoB tool, potential improvements have been identified. A formal evaluation found some bias domains to be confusing at times, with assessment of bias due to incomplete outcome data and selective reporting of outcomes causing particular difficulties and confusion over whether studies that were not blinded should automatically be considered to be at high risk of bias. More guidance on incorporating risk-of-bias assessments into meta-analyses and review conclusions is also needed. A review of comments and user practice found that both Cochrane and non-Cochrane systematic reviews often implemented the RoB tool in non-standard ways. Few trials are assessed as at low risk of bias, and judgments of unclear risk of bias are common. Empirical studies have found only moderate reliability of risk-of-bias judgments. (p. 1)

Our findings demonstrate that considerable care is needed in systematic reviews in the process of bias assessment but also that each reader must make their own judgments on the studies included in reviews and the conclusions reached.

Why are the reviews so uncritical of the Jeong et al. trial?

This is the most important question to address, and it goes to the heart of our critique of original research and evidence reviews in the field of creative arts therapies and arts activities and health.[28,30,64,65] The question can be applied to all the reviews we consider. In the poorer reviews, it is difficult not to draw the conclusion that the authors were lax in summarizing the findings of the study or simply took on trust the conclusions reached by Jeong et al.[32] The well-conducted systematic reviews are more careful in their evaluation of the Jeong et al. study[32] but are guided by quality scales (e.g., the Jadad scale[62] in Strassel et al.[51]) and the Cochrane RoB tool. None of the reviews we consider offer a more radical critique and question the "mild depression" label applied to the girls, the lack of reference to clinical cutoff points and minimal clinically important change scores, the conduct of the study in one school with the use of a waiting list control, and the lack of details on the assaying of neurohormones.

An important factor here, we believe, is that the increasingly standardized procedures for the appropriate conduct of systematic reviews and meta-analysis, and the use of the Cochrane RoB Tool,[11–13] appear to direct reviewers' attention toward design issues and render them blind to other aspects of primary studies which require scrutiny.

Our findings are in line with the critiques offered by MacLure[20] and Greenhalgh et al.[21] MacLure[20] offers a trenchant analysis of the difficulties associated with systematic review methodology and the way in which it disciplines both reading and writing within tight constraints. The approach, she claims:

> "… degrades the status of reading and writing as scholarly activities tend to result in reviews with limited capacity to inform policy or practice and constitutes a threat to quality and critiques in scholarship and research" (p. 393).

A similar point is made by Greenhalgh et al.[21] in questioning the widespread assumption that systematic reviews are superior to narrative reviews. Narrative reviews, they argue, are concerned to present

"an authoritative argument based on informed wisdom that is convincing to an audience of fellow experts" (p. 3). This perspective is relevant to the best of the systematic reviews and meta-analyses on dance therapy research considered here,[47–50] which provide interesting and insightful introductions on the development and practice of dance therapy and elucidate the theoretical mechanisms that account for how dance may achieve therapeutic change.

CONCLUSIONS

There are limitations to the work we report here. We have only undertaken an analysis of one target paper by Jeong et al.[32] and considered the way it is treated in 14 reviews/meta-analyses. However, two earlier papers[28,30] have revealed the same concerns, and our work, therefore, leads to the following recommendations:

- Further studies following the innovative method demonstrated in this and previous papers[28,30] are needed to assess the accuracy and credibility of systematic reviews in the field of arts and health.
- Systematic reviews should be properly focused, preregistered in PROSPERO,[66] and conducted to a high standard following current PRISMA guidelines.[14,15] Particular attention is needed to double-check judgments of bias.[11–13]
- 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. The time needed to undertake a satisfactory peer review may be considerably longer than most prospective journal reviewers are prepared to commit.
- Greater attention is needed in the field of arts and health to replicate 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. It is a matter of serious concern that the trial conducted by Jeong et al.[32] has never been replicated.
- RCTs have an important role to play in evaluating creative arts therapies and arts for health programs, but qualitative studies are essential too[68,70,71] It should be recognized, however, that neither participants nor professionals facilitating arts activities can be blind to the activity they are engaged in and that all such studies are vulnerable to expectation bias.
- Systematic reviews and meta-analyses should be conducted by an interdisciplinary team covering relevant subject matter and quantitative expertise.
- Practitioners and researchers in the wider field of arts and health should approach evidence reviews, systematic reviews, and meta-analyses with an appropriate degree of caution.[67]

AUTHOR CONTRIBUTIONS

K.G.-H.: Conceptualization, funding acquisition, investigation, methodology, project administration, supervision, validation, writing (original draft, review, and editing); L.T.-H.: Funding acquisition, methodology, writing (review and editing); A.K.S.-W.: Investigation, writing (review and editing); A.C.B.: Investigation, formal analysis, methodology, writing (review and editing); S.C.: Conceptualization, formal analysis, investigation, methodology, visualization, writing (original draft, review, and editing). All authors approved the final submission.

ACKNOWLEDGMENTS

We are grateful to Urs M. Nater, Matt McCrary, and Helen Payne for helpful comments on earlier drafts of this paper. K.G.-H. and A.K.S.-W. were supported by Salzburg Land, Project "Art is a doctor." The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

ORCID

Katarzyna Grebosz-Haring https://orcid.org/0000-0002-2922-5129

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/nyas.15006

REFERENCES

1. 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
2. Fancourt, D., & Finn, S. (2019). *What is the evidence on the role of the arts in improving health and wellbeing? A scoping review*. WHO Regional Office for Europe.
3. Fancourt, D., Aughterson, H., & Warren, K. (2020). *Evidence summary for policy: The role of arts in improving health and wellbeing*. University College London.
4. 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*(12), e01014. https://doi.org/10.1016/j.heliyon.2018.e01014
5. Grebosz-Haring, K., Schuchter-Wiegand, A. K., Feneberg, A. C., Skoluda, N., Nater, U. M., Schütz, S., & Thun-Hohenstein, L. (2021). 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
6. 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.
7. Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., R Pitkin, D Rennie, K F Schulz, D Simel, & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA*, *276*(8), 637–639. https://doi.org/10.1001/jama.276.8.637
8. Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P. C., & Lang, T. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, *134*(8), 663–694. https://doi.org/10.7326/0003-4819-134-8-200104170-00012
9. Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010).

CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c869. https://doi.org/10.1136/bmj.c869
10. Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials*, *16*(1), 62–73. https://doi.org/10.1016/0197-2456(94)00031-w
11. Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928. https://doi.org/10.1136/bmj.d5928
12. Turner, L., Boutron, I., Hróbjartsson, A., Altman, D. G., & Moher, D. (2013). The evolution of assessing bias in Cochrane systematic reviews of interventions: Celebrating methodological contributions of the Cochrane Collaboration. *Systematic Reviews*, *2*(1), 79. https://doi.org/10.1186/2046-4053-2-79
13. Sterne, J. A. C., 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., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898
14. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71
15. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535
16. Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(1), 10. https://doi.org/10.1186/1471-2288-7-10
17. 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. *BMJ*, *358*, j4008. https://doi.org/10.1136/bmj.j4008
18. Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *350*, g7647. https://doi.org/10.1136/bmj.g7647
19. Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003
20. MacLure, M. (2005). 'Clarity bordering on stupidity': Where's the quality in systematic review? *Journal of Education Policy*, *20*(4), 393–416. https://doi.org/10.1080/02680930500131801
21. Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, *48*(6), e12931. https://doi.org/10.1111/eci.12931
22. Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210
23. 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*(4), 518–520. https://doi.org/10.1007/s00134-017-5039-y
24. Eysenck, H. J. (1994). Meta-analysis and its problems. *BMJ*, *309*(6957), 789–792. https://doi.org/10.1136/bmj.309.6957.789
25. Eysenck, H. J. (1995). Meta-analysis or best-evidence synthesis? *Journal of Evaluation in Clinical Practice*, *1*(1), 29–36. https://doi.org/10.1111/j.13652753.1995.tb00005.x
26. 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
27. Sharpe, D., & Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology /Psychologie Canadienne*, *61*, 377–387. https://doi.org/10.1037/cap0000215
28. Grebosz-Haring, K., Thun-Hohenstein, L., Schuchter-Wiegand, A. K., Irons, Y., Bathke, A., Phillips, K., & 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
29. 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.
30. Clift, S., Grebosz-Haring, K., Thun-Hohenstein, L., Schuchter-Wiegand, A. K., & Bathke, A. (2022). The need for robust critique of arts and health research: An examination of the Goldbeck and Ellerkamp (2012) RCT of music therapy for anxiety in children, and its treatment in four systematic reviews. *Approaches: An Interdisciplinary Journal of Music Therapy*, Advance online publication, https://approaches.gr/clift-a20220811/
31. 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
32. Jeong, Y.-J., Hong, S.-C., Lee, M. S., Park, M.-C., Kim, Y.-K., & Suh, C.-M. (2005). Dance movement therapy improves emotional responses and modulates neurohormones in adolescents with mild depression. *International Journal of Neuroscience*, *115*(12), 1711–1720. https://doi.org/10.1080/00207450590958574
33. 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, https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021193283
34. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004
35. Derogatis, L. R. (1994). *Symptom Checklist-90-Revised (SCL-90-R)*. https://www.pearsonclinical.co.uk/
36. Pascoe, M. C., Bailey, A. P., Craike, M., Carter, T., Patten, R., Stepto, N. K., & Parker, A. G. (2020). Exercise interventions for mental disorders in young people: A scoping review. *BMJ Open Sport & Exercise Medicine*, *6*(1), e000678. https://doi.org/10.1136/bmjsem-2019-000677
37. Pascoe, M. C., Bailey, A. P., Craike, M., Carter, T., Patten, R., Stepto, N. K., & Parker, A. (2020). Poor reporting of physical activity and exercise interventions in youth mental health trials: A brief report. *Early Intervention in Psychiatry*, *15*(5), 1414–1422. https://doi.org/10.1111/eip.13045
38. Schulz, K. F., Altman, D. G., & Moher, D., CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*(1), 18. https://doi.org/10.1186/1741-7015-8-18
39. Cunningham, J. A., Kypri, K., & McCambridge, J. (2013). Exploratory randomized controlled trial evaluating the impact of a waiting list control design. *BMC Medical Research Methodology*, *13*(1), 150. https://doi.org/10.1186/1471-2288-13-150

40. Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., Jensen-Doss, A., Hawley, K. M., Krumholz Marchette, L. S., Chu, B. C., Weersing, V. R., & Fordwood, S. R. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist*, *72*(2), 79–117. https://doi.org/10.1037/a0040360
41. Derogatis, L. R. (1977). *SCL-90-R: Administration, scoring and procedures manual*. Clinical Psychometric Research.
42. Preti, A., Carta, M. G., & Petretto, D. R. (2019). Factor structure models of the SCL 90-R: Replicability across community samples of adolescents. *Psychiatry Research*, *272*, 491–498. https://doi.org/10.1016/j.psychres.2018.12.146
43. Stoffel, M., Neubauer, A. B., & Ditzen, B. (2021). How to assess and interpret everyday life salivary cortisol measures: A tutorial on practical and statistical considerations. *Psychoneuroendocrinology*, *133*, 105391. https://doi.org/10.1016/j.psyneuen.2021.105391
44. Strahler, J., Skoluda, N., Kappert, M. B., & Nater, U. M. (2017). Simultaneous measurement of salivary cortisol and alpha-amylase: Application and recommendations. *Neuroscience and Biobehavioral Reviews*, *83*, 657–677. https://doi.org/10.1016/j.neubiorev.2017.08.015
45. McCrary, J. M., & Altenmüller, E. (2021). Mechanisms of music impact: Autonomic tone and the physical activity roadmap to advancing understanding and evidence-based policy. *Frontiers in Psychology*, *12*, 727231. https://doi.org/10.3389/fpsyg.2021.727231
46. Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x
47. Koch, S., Kunz, T., Lykou, S., & Cruz, R. (2014). Effects of dance movement therapy and dance on health-related psychological outcomes: A meta-analysis. *Arts in Psychotherapy*, *41*(1), 46–64. https://doi.org/10.1016/j.aip.2013.10.004
48. Koch, S. C., Riege, R. F. F., Tisborn, K., Biondo, J., Martin, L., & Beelmann, A. (2019). Effects of dance movement therapy and dance on health-related psychological outcomes. A meta-analysis update. *Frontiers in Psychology*, *10*, 1806. https://doi.org/10.3389/fpsyg.2019.01806
49. Meekums, B., Karkou, V., & Nelson, E. A. (2015). Dance movement therapy for depression. *Cochrane Database of Systematic Reviews (Online)*, *2015*(2), Cd009895. https://doi.org/10.1002/14651858.CD009895.pub2
50. Karkou, V., Aithal, S., Zubala, A., & Meekums, B. (2019). Effectiveness of dance movement therapy in the treatment of adults with depression: A systematic review with meta-analyses. *Frontiers in Psychology*, *10*, 936. https://doi.org/10.3389/fpsyg.2019.00936
51. Strassel, J. K., Cherkin, D. C., Steuten, L., Sherman, K. J., & Vrijhoef, H. J. (2011). A systematic review of the evidence for the effectiveness of dance therapy. *Alternative Therapies in Health and Medicine*, *17*(3), 50–59.
52. Kiepe, M.-S., Stöckigt, B., & Keil, T. (2012). Effects of dance therapy and ballroom dances on physical and mental illnesses: A systematic review. *Arts in Psychotherapy*, *39*(5), 404–411. https://doi.org/10.1016/j.aip.2012.06.001
53. Millman, L. S. M., Terhune, D., Hunter, E., & Orgs, G. (2021). Towards a neurocognitive approach to dance movement therapy for mental health: A systematic review. *Clinical Psychology & Psychotherapy*, *28*(1), 24–38. https://doi.org/10.1002/cpp.2490
54. Tao, D., Gao, Y., Cole, A., Baker, J., Gu, Y., Supriya, R., Tong, T. K., Hu, Q., & Awan-Scully, R. (2022). The physiological and psychological benefits of dance and its effects on children and adolescents: A systematic review. *Frontiers in Physiology*, *13*, 925958. https://doi.org/10.3389/fphys.2022.925958
55. Mala, A., Karkou, V., & Meekums, B. (2012). Dance/movement therapy (D/MT) for depression: A scoping review. *Arts in Psychotherapy*, *39*(4), 287–295. https://doi.org/10.1016/j.aip.2012.04.002
56. Calçada, J., & Gilham, C. (2022). Biodanza and other dance forms as a vehicle for social emotional learning in schools: A scoping review. *LEARNing Landscapes*, *15*(1), 53–73. https://doi.org/10.36510/learnland.v15i1.1059
57. Grudzińska, A., & Izdebski, P. (2018). The effect of dance therapy on patients with mental and somatic disorders - A review of research. *Rehabilitacja Medyczna*, *22*, 32–37. DOI: https://doi.org/10.5604/01.3001.0012.0896
58. Lopez-Nieves, I., & Jakobsche, C. E. (2022). Biomolecular effects of dance and dance/movement therapy: A review. *American Journal of Dance Therapy*, *44*(2), 241–263. https://doi.org/10.1007/s10465-022-09368-z
59. Lossing, A., Moore, M., & Zuhl, M. (2017). Dance as a treatment for neurological disorders. *Body, Movement and Dance in Psychotherapy*, *12*(3), 170–184. doi:10.1080/17432979.2016.1260055
60. Lykesas, G., Chatzopoulos, D., Neratzoula, V., Nikolaki, E., Douka, S., & Bakirtzoglou, P. (2022). Reviewing available online publications on the effect of dance on the physical and mental health of children and adolescents. *Central European Journal of Sport Sciences and Medicine*, *3*(39), 17–26. https://doi.org/10.18276/cej.2022.3-02
61. Meekums, B., Karkou, V., & Nelson, E. A. (2012). Dance movement therapy for depression. *Cochrane Database of Systematic Reviews*, *6*, CD009895. https://doi.org/10.1002/14651858.CD009895
62. Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, *17*(1), 1–12. https://doi.org/10.1016/01972456(95)00134-4
63. Siontis, K. C., & Ioannidis, J. P. A. (2018). Replication, duplication, and waste in a quarter million systematic reviews and meta-analyses. *Circulation: Cardiovascular Quality and Outcomes*, *11*(12), e005212. https://doi.org/10.1161/CIRCOUTCOMES.118.005212
64. Habibi, A., Kreutz, G., Russo, F., & Tervaniemi, M. (2022). Music-based interventions in community settings: Navigating the tension between rigor and ecological validity. *Annals of the New York Academy of Sciences*, *1518*(1), 47–57. https://doi.org/10.1111/nyas.14908
65. Grau-Sánchez, J., Jamey, K., Paraskevopoulos, E., Dalla Bella, S., Gold, C., Schlaug, G., Belleville, S., Rodríguez-Fornells, A., Hackney, M. E., & Särkämö, T. (2022). Putting music to trial: Consensus on key methodological challenges investigating music-based rehabilitation. *Annals of the New York Academy of Sciences*, *1518*(1), 12–24. https://doi.org/10.1111/nyas.14892
66. Page, M. J., Shamseer, L., & Tricco, A. C. (2018). Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Systematic Reviews*, *7*(1), 32. https://doi.org/10.1186/s13643-018-0699-4
67. Munn, Z., Pollock, D., Barker, T. H., Stone, J., Stern, C., Aromataris, E., Schünemann, H. J., Clyne, B., Khalil, H., Mustafa, R. A., Godfrey, C., Booth, A., Tricco, A. C., & Pearson, A. C. (2022). The Pandora's Box of Evidence Synthesis and the case for a living Evidence Synthesis Taxonomy. *BMJ Evidence-Based Medicine*, Advance online publication, https://ebm.bmj.com/content/early/2022/10/14/bmjebm-2022-112065

How to cite this article: Grebosz-Haring, K., Thun-Hohenstein, L., Schuchter-Wiegand, A. K., Bathke, A. C., & Clift, S. (2023). The need for robust critique of arts and health research: Dance-movement therapy, girls, and depression. *Ann NY Acad Sci.*, 1–12. https://doi.org/10.1111/nyas.15006