

## The need for robust critique of arts and health research: the treatment of the Gene Cohen et al. (2006) paper on singing, wellbeing and health in subsequent evidence reviews

Stephen Clift, Katarzyna Grebosz-Haring, Leonhard Thun-Hohenstein, Anna Katharina Schuchter-Wiegand, Arne Bathke & Mette Kaasgaard

To cite this article: Stephen Clift, Katarzyna Grebosz-Haring, Leonhard Thun-Hohenstein, Anna Katharina Schuchter-Wiegand, Arne Bathke & Mette Kaasgaard (05 Jan 2024): The need for robust critique of arts and health research: the treatment of the Gene Cohen et al. (2006) paper on singing, wellbeing and health in subsequent evidence reviews, Arts & Health, DOI: [10.1080/17533015.2023.2290075](https://doi.org/10.1080/17533015.2023.2290075)

To link to this article: <https://doi.org/10.1080/17533015.2023.2290075>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 05 Jan 2024.



[Submit your article to this journal](#)



Article views: 77



[View related articles](#)



[View Crossmark data](#)

# The need for robust critique of arts and health research: the treatment of the Gene Cohen et al. (2006) paper on singing, wellbeing and health in subsequent evidence reviews

Stephen Clift<sup>a</sup>, Katarzyna Grebosz-Haring<sup>b</sup>, Leonhard Thun-Hohenstein<sup>c</sup>,  
Anna Katharina Schuchter-Wiegand<sup>a</sup>, Arne Bathke<sup>d</sup> and Mette Kaasgaard<sup>e</sup>

<sup>a</sup>Sidney De Haan Research Centre for Arts and Health, Canterbury Christ Church University, Canterbury, UK;

<sup>b</sup>Grebosz-Haring Department of Art History, Musicology and Dance Studies, Paris Lodron University, Salzburg/University Mozarteum, Salzburg, Austria; <sup>c</sup>Grebosz-Haring Department of Art History, Musicology and Dance Studies, Paracelsus Medical University, Salzburg, Austria; <sup>d</sup>Grebosz-Haring Department of Art History, Musicology and Dance Studies, Paris Lodron University, Salzburg, Austria; <sup>e</sup>Institute of Regional Health Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

## ABSTRACT

**Background:** This paper considers weaknesses in a study by Cohen et al. (2006) on the impacts of community singing on health. These include high demand characteristics, lack of attention to attrition, flawed statistical analysis, and measurement. Nevertheless, the study is uncritically cited, in evidence reviews, with findings taken at face value.

**Methods:** Google Scholar, SCOPUS and BASE citation functions for Cohen et al. identified 32 evidence reviews in peer-reviewed journals. Eleven of these reviews, published between 2010 and 2023, focused on creative arts interventions.

**Results:** We demonstrate limitations in the Cohen et al. research which undermine the conclusions they reach regarding the health benefits of group singing. Subsequent evidence reviews take the findings at face value and offer little critical commentary.

**Discussion:** We consider what is needed to improve evidence reviews in the field of creative arts and health research.

**Conclusions:** A more robust approach is needed in reviewing research evidence in the field of arts and health. The Cohen et al. paper is not suitable for inclusion in future evidence reviews.

## ARTICLE HISTORY

Received 29 June 2023

Accepted 27 November 2023


## KEYWORDS

Singing; creative arts; older people; health; evidence reviews

## Introduction

In three previous publications on the need for robust critique of research in arts and health, we have focused on examples of controlled trials on creative arts therapies for children and young people with mental health challenges. The first of these three publications considered a research study on art therapy and its treatment in three systematic reviews (removed to anonymise). The second focused on a music therapy trial and its treatment in two systematic

**CONTACT** Stephen Clift  [stephen.clift@canterbury.ac.uk](mailto:stephen.clift@canterbury.ac.uk)  Sidney De Haan Research Centre for Arts and Health, Canterbury Christ Church University, Canterbury, UK

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/17533015.2023.2290075>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

reviews and two meta-analyses (removed to anonymise), and the third looked at the treatment of a controlled trial on dance-movement therapy in no less than 14 reviews including narrative reviews, scoping reviews, systematic reviews, meta-analyses and a Cochrane review (removed to anonymise). The conclusions drawn in each paper were the same: authors of the evidence reviews considered accepted results in the target paper at face value and generally failed to recognise serious limitations and weaknesses in the research. Furthermore, peer reviewers and journal editors prior to publication of the evidence reviews must also have failed to recognise the lack of criticality in the reviews. We have also reported evidence of subjectivity in risk of bias assessments using the Cochrane Risk of Bias Tool of the same paper across even meticulously conducted systematic reviews (Clift et al., 2022; Gebosz-Haring et al., 2023). Furthermore, we found errors in data extraction for the purposes of meta-analyses, despite judgements being made independently by two reviewers (Clift et al., 2022). We now turn our attention to the wider field of arts and health research, to consider the way in which a specific example of research is treated in evidence reviews, including systematic reviews and meta-analyses.

In the present paper, we take as the focus the “ground-breaking” study (Cutler, 2019), by Cohen et al (2006, 2007), on the impacts of community singing for older people on physical and mental health. We consider the ways in which this research is presented in 11 subsequent reviews of research concerned with the health and wellbeing benefits of engagements in creative arts, music and singing. It is particularly appropriate to consider research on singing and health, since several of the earliest studies in the development of research in “arts and health” were concerned with the possible value of community singing for health and wellbeing (e.g. Bailey & Davidson, 2002; Beck et al., 2000; Clift & Hancox, 2001; Kreutz et al., 2004). The Cohen et al.’s study is also recently very widely cited and is seen as a precursor to later quasi-experimental (Maury & Rickard, 2022) and randomised controlled trials on the value of singing for the wellbeing of older people (Coulton et al., 2015; Johnson et al., 2020).

As with our earlier papers (Clift et al., 2022; Grebosz-Haring et al., 2022; Grebosz-Haring, et al., 2023), our approach is informed by a critique (Clift, et al., 2021) of two major scoping reviews of the arts and health research literature for their lack of critical scrutiny of studies included (Fancourt & Finn, 2019; Fancourt et al., 2020). We also take account of recent critical perspectives on the conduct of systematic reviews and meta-analyses in medicine, health care, and education.

We recognise that evidence reviews are conducted for different purposes and can take a variety of forms (Grant & Booth, 2009), and even systematic reviews may serve a variety of functions (Munn et al., 2018). Whatever approach is adopted, reviews may have both advantages and disadvantages, and can vary in quality. If the concern, however, is to assess the effectiveness of interventions with respect to specified outcomes, systematic reviews and meta-analyses are commonly considered to be the “gold standard” at the top of “evidence hierarchies” and Stegenga (2011) has suggested that meta-analyses may be considered as the “platinum standard” in the synthesizing of evidence.

Detailed guidance is currently available on the conduct of systematic reviews (Aromataris & Munn, 2020; Higgins et al., 2023), but even so, a number of critics have raised substantial reservations regarding the practice of systematic reviews and meta-analysis and identified weaknesses in their execution. MacLure (2005), for example, in an early paper with continued relevance is critical of systematic reviews on educational

topics conducted over the period 2002–4 for their lack of close reading of primary studies. Greenhalgh et al. (2018) acknowledge that narrative reviews may have limitations but question the view that systematic reviews are necessarily superior to such reviews. Ioannidis (2016) suggests that the production of systematic reviews and meta-analyses “has reached epidemic proportions” and he considers most systematic reviews and meta-analyses as “unnecessary, misleading, and/or conflicted” (p.468). Møller et al. (2018) go further and argue that many systematic reviews ‘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). Moore et al. (2023) point out that systematic reviews can be rendered valueless by inclusion of poor quality and under-powered primary studies. Stegenga (2011) and Murad et al. (2016) also highlight that meta-analyses can be compromised by subjective judgements at every stage of their execution.

### ***The Cohen et al. (2006) paper on singing***

The Cohen et al. (2006) paper and a subsequent follow-up report in 2007, were subject to careful critical scrutiny shortly after their appearance by (Clift, et al., 2008) an unpublished mapping of singing for wellbeing research. In this, they identified substantial limitations in the 2006 report and further concerns in the 2007 follow-up. Subsequently, updated and published mapping of singing research (Clift, et al., 2010) stated of the Cohen study: “... the study has methodological and analytical weaknesses (...) which mean that the authors’ conclusions have to be interpreted cautiously’ (p.5). Reservations regarding the weaknesses of the Cohen et al.’s (2006, 2007) study were also expressed by (Clift, et al., 2016; Clift, 2020).

Our current concerns about the lack of critical treatment of Cohen et al. (2006) in evidence reviews were raised by the account of this study appearing in a report from the Baring Foundation on creative aging (Cutler, 2019). Cutler argues that involvement in the arts is beneficial for older people and refers to the Cohen et al. study as a “landmark study” in this field:

Published in *The Gerontologist* (Vol 46, no 6, pages 726–734), the study followed 300 subjects with a median age of 80. One group was involved in arts programmes and the control group was not. The study suggested that involvement in the arts led to better health, fewer visits from doctors, less medication, increased physical activity and social engagement. This led to the claim that such programmes could result in a reduction of \$6.3 billion dollars at that time to the US public purse. (p.8)

However, this account of the study is partial and inaccurate. No mention is made of the fact that this paper reports an evaluation of a community singing programme, and no indication is given of significant limitations and weaknesses in the design and methods of this research.

More recently, the Cohen et al. (2006) study is referenced without critical scrutiny in two substantial grey literature reports. Hallam and Himonides (2022), in a wide-ranging review of evidence on the “The Power of Music”, give the following account:

The choir group reported a higher overall rating of physical health, fewer visits to the doctor, less medication use, fewer falls, and fewer other health problems in comparison with a control group, who had carried on with their usual activities and did not participate in the choir. There was evidence of higher morale, a reduction in loneliness and increased activity, while the comparison group experienced a significant decline in activities. Cohen and colleagues argued that sense of control, as well as social engagement, were the most likely mechanisms responsible for the positive outcomes. (p.488)

Furthermore, Bone and Fancourt (2022) in a review of research on “Arts, Culture & the Brain” refer to the study by Cohen et al. (2006) among others, in support of the claim that

Reviews have shown that participatory arts interventions can reduce loneliness and social isolation in older adults. For example, choir groups, making music, and arts and crafts programmes can decrease loneliness, facilitate new social relationships, and increase perceptions of closeness among participants. (p.7)

The quotations given above from Cutler (2019) and Hallam and Himonides (2022) provide an account of the findings from the Cohen et al. study. However, before describing the method employed in this exercise in robust critique, it is important to provide our own non-evaluative summary of the methodology and findings from the Cohen research.

The Cohen et al. study took place in Washington DC and had a two-group comparative design. Participants were “166 healthy, ambulatory older adults from the Washington, DC, area” (p.726) who were recruited separately to a community singing group and to the usual activity comparison group. Thus, as Cohen et al. acknowledge, the study is “quasi-experimental” as participants were not randomly assigned to the two arms of the study.

Ten measures were used at baseline and after one year to assess a range of health-related variables: a subjective overall health rating, number of doctor visits, number of over-the-counter medicines, number of falls, and number of health problems, all in the previous 12 months. In addition, participants completed three structured questionnaires to measure morale (Lawton, 1975), depression (Yesavage and Sheikh 1986) and loneliness (Russell, 1996). They were also asked to report on their weekly activities and total number of activities over the previous year.

In analysing the findings, Cohen et al. first compared the two groups at baseline, and established that for most variables the groups were equivalent, except that statistically significant differences were found for a number of health problems, and levels of depression and loneliness in favour of the singing group. Cohen assessed the impact of a year of regular singing by comparing the two groups on the measures at follow-up. Where no differences were found at baseline, simple comparisons were undertaken with either independent t-tests or chi-square tests, but where the groups differed at the outset, Analysis of Covariance (ANCOVA) was used to take account of baseline differences. For depression and the total social activity measure, no differences were found at follow-up and Cohen reports post-hoc within-group comparisons using paired t-tests.

Table 1 reports the results for singing and comparison group after a year based on Table 2 in the Cohen et al. (2006) paper (p. 729) and details of the statistical analyses undertaken described on pp. 730–31. The first five measures relate to physical health, the next three relate to mental wellbeing and the last two assess how active participants were. Cohen et al. (2006) use a liberal 10% criterion for

**Table 1.** Means (SD) and statistical analysis reported by Cohen et al. (2006) on 10 measures at 12-months follow-up.

Variable	Singing group (n = 77)	Control group (n = 64)	Statistical test and p-values
<b>Health indicator</b>			
Overall health rating***	7.97 (1.58)	7.25 (1.91)	t = -2.39, p = 0.01 <sup>a</sup>
No. Doctor visits**	6.73 (7.00)	10.84 (14.49)	t = 2.06, p = 0.04 <sup>b</sup>
No. over-counter medicines***	0.23 (0.69)	0.55 (1.30)	F = 10.02, p < 0.01
No. falls**	0.23 (0.69)	0.55 (1.30)	t = 1.82, p = 0.05 <sup>c</sup>
Other health problems*	0.40 (0.46)	0.45 (0.50)	$\chi^2 = 3.58, p = 0.06^d$
<b>Mood indicator</b>			
Morale**	14.08 (2.66)	13.06 (3.29)	t = -1.92, p < 0.06 <sup>e</sup>
Depression	1.14 (1.84)	1.84 (1.89)	F not reported, n.s.
Loneliness*	34.60 (7.86)	37.02 (10.33)	F = 3.08, p = 0.08
<b>Level of activity</b>			
No. weekly activities**	4.29 (2.55)	2.58 (1.82)	t = -4.62, p < 0.01
No. total activities <sup>f</sup>	10.55 (5.04)	8.05 (3.70)	t not reported, n.s.

Based on Table 2 in Cohen et al. (2006, p. 729) with details of statistical tests given in the text (pp.730–1).

\* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ , as given in Table 2.

For over-counter medicines and measures of depression and loneliness, significant differences were found at baseline and Analysis of Covariance was used to take account of initial differences.

<sup>a</sup>Critical 2-tailed value of t for  $p = 0.01$ , is 2.63, and critical 1-tailed value of t for  $p = 0.01$ , is 2.36. Clear, therefore, that change in overall health rating is judged using a 1-tailed test.

<sup>b</sup>Unclear why the  $p$  value is reported at 0.04 in the text, when the t-value is above the  $p = 0.05$  critical value of 1.98 for a 2-tailed test, and clearly above the critical value of 1.66 for a 1-tailed test.

<sup>c</sup>Critical 2-tailed value of t for  $p = 0.05$ , is 1.98, and critical 1-tailed value of t for  $p = 0.05$ , is 1.66. Clear, therefore, that change in number of falls is judged using a 1-tailed test.

<sup>d</sup>Critical value of  $\chi^2$  (df = 1) for  $p = 0.05$  is 3.84, which appears to account for the reported  $p$  value of 0.06.

<sup>e</sup>Value for  $p < 0.06$  is reported in the text, but two asterisks (\*\*) are given next to Morale in column 1. The reported t-value is above the critical 1-tailed value, however, and so using a 1-tailed test the difference is significant at 5%. It seems likely, therefore, that the reported value  $p < 0.06$  is a typographical error.

<sup>f</sup>No difference found in total number of activities at follow up. Post-hoc paired t-tests showed no change for the singing group from baseline, but a significant decline for the comparison group.

**Table 2.** Cohen et al. (2006) post-test means on four measures (values rounded to the nearest whole number).

	Singing group (n = 77)	Comparison group (n = 64)
General health (Scale 0–10, higher score = better health)	8	7
Morale (Geriatric Morale Scale 0–17, 9+ = good morale)	14	13
Depression (Geriatric Depression Scale 0–15, 5+ = depression)	1	2
Loneliness (UCLA loneliness Scale, 20–80, 50+ = loneliness)	35	37

identifying significant changes on the grounds that the study was “exploratory.” In addition, it is clear from considering the reported t-values that a one-tailed criterion was used for rejecting the null hypothesis. Chi-square and ANCOVA are non-directional tests of the null hypothesis, and thus two-tailed (Gorard, 2021). It should also be noted that non-significant results for the use of “prescription medicines” is not reported in their tables, but details can be found in the body of the text (p.720).

## Method

In this paper, we followed a modified version of the method developed in our previous robust review papers (removed to anonymise).

- (1) As a basis for a critical account of the Cohen et al. study, four authors with PhDs in Statistics, Child Psychiatry, Musicology and Singing in Pulmonary Rehabilitation, conducted independent critical appraisals of the Cohen et al. study drawing on their knowledge of research. This exercise was undertaken without awareness of earlier critical commentaries on the Cohen et al. research by the first author (removed to anonymise). Assessments by the four authors were shared and key limitations of the research were agreed among all authors. In addition, the first author and two others used the Joanna Briggs Institute (JBI) appraisal tool for quasi-experimental studies (see <https://jbi.global/critical-appraisal-tools>) to assess the Cohen et al. study independently. This tool consists of nine questions, which are answered “yes”, “no” or ‘unclear (e.g. “4. Was there a control group?” and “8. Were outcomes measured in a reliable way?”). Differences in the three authors' judgements were resolved through discussion.
- (2) In order to identify subsequent reviews citing Cohen, we used Google Scholar as in our previous papers. In addition, however, for the current exercise we also undertook citation searches using Scopus and the Bielefeld Academic Search Engine (BASE), to check on the completeness of the evidence reviews identified through Google Scholar (Gusenbauer, 2019; Gusenbauer and Haddaway, 2020; Martinmartín et al., 2021). Each identified art-focused review located through our search strategy was analysed to extract details of the Cohen et al. results referred to, and any critical comments on the methodology employed, the findings and conclusions drawn.

## Results

### *A robust critique of the Cohen et al. study*

A detailed critique of the Cohen paper was previously provided by (removed to anonymise). In the present paper, however, we focus specifically on four main issues, and argue that the Cohen et al. study (Cohen et al., 2006) provides no convincing scientific evidence for the health benefits of group singing for older people.

Firstly, all studies of singing and health are vulnerable to a range of possible biases since participants and facilitator cannot be blind to the activity and its purpose. This is certainly a serious consideration for the Cohen study given the information provided for people joining the singing group. Here, the details of how participants were recruited and show that “demand characteristics” (Nicols & Maner, 2008) built into the study were substantial:

The notice for the intervention group . . . sought singers for a chorale; no singing experience was required, and the study's purpose was to explore the impact of this activity on general health and mental health as well as involvement in overall individual and group activities of older adults living in the community. (p.728)

Participants were primed to expect benefits, and the measures used are all transparent. Even small consistent biases, such as “expectancy” and “cooperation” could translate into “significant” effects in comparing singers and controls.

Secondly, there is a failure to consider the impact of sample attrition on the post-test results. The analysis at baseline involved the total sample of people recruited into the study ( $N = 166$ ), whereas at 12-month follow-up, data are presented for 141 participants, which represents a loss of 25 participants (i.e. a 15% attrition rate, most likely for participants with poorer health). What Cohen et al. should have done instead was to report the baseline results for the 141 participants followed up. They should also have reported a comparison of baseline characteristics of participants who did not complete the study. As it is, we cannot rule out the possibility that differences claimed by Cohen et al. at follow-up do not reflect existing differences at baseline for the reduced sample. This is particularly true for those variables in which a simple two-group comparison was made at follow-up between the singing and control group.

Thirdly, there are flaws in the statistical treatment of the results. As all of the measures employed are nominal or ordinal in character, non-parametric tests would have been a more appropriate choice for inference (as Cohen et al. have done when applying chi-square tests to the health problems data). However, it should be noted that non-parametric inference methods are usually not based on the same effect measures as parametric tests which need to be considered when interpreting the results.

A further problem, however, is that for the independent t-tests, it is clear that Cohen et al. employed a one-tailed criterion for judging statistical significance (removed to anonymise, for full details, and notes a, c and e in [Table 1](#)). Directional tests are inappropriate given the need for “equipose” in studies evaluating a health intervention. In addition, for a non-randomised, exploratory study of this kind, two-tailed tests would have been more appropriate. If this more stringent standard is applied, the reported one-sided p-values should be doubled. In addition, the use of a 10% criterion for judging significance is also too liberal and inappropriate given the nature of the study and the measures employed and no corrections are made for the fact that multiple statistical tests are applied (Ellis, 2010).

Overall, a more cautious approach to inferential statistical analysis would have been appropriate. Indeed, considering the design limitations, it may have been preferable to focus on descriptive reporting, and only use (nonparametric) inferential methods for some of the key comparisons, and with appropriate caution in interpreting any significant findings.

Given these considerations, the number of apparently significant results is considerably reduced, with only the difference for ratings of general health, number of over-the-counter medications and regular weekly activities favouring the singing group. With respect to the later measure, however, it is not clear from Cohen et al.’s account whether members of the singing group were asked to exclude the singing activity. If not, the difference at follow-up may simply reflect the fact that compared with the control group who reported activities as usual, they were also singing each week.

Finally, and most critical of all, there are substantial limitations in the nature and interpretation of the general health rating, and the measures used as “mood” indicators. Each of these measures, as noted earlier, is ordinal in character, and consequently, non-parametric techniques for statistical analysis would have been more appropriate than the



parametric tests used. In addition, given the ordinal level of measurement, reporting mean values to two decimal places represents an inappropriate level of precision. Table 2 reports means for these measures rounded to the nearest whole number as a reasonable approximation. We now need to consider the psychometric properties of the scales themselves as this has an important bearing on how the values are interpreted.

For the “general health” variable, participants were asked to give a subjective rating of health from 0 to 10, with 0 meaning “worst” and 10 meaning “best”. As we can see, participants rate their health highly, and the difference is a mere one point on this scale. It is impossible to say what this difference means regarding the health of the two groups.

For the morale, depression and loneliness data, we need to do some additional research to find further information on the scales used, as specific details are not given by Cohen et al. (2006).

For morale, the scale used is the Philadelphia Geriatric Center Morale Scale (Lawton, 1975). This questionnaire consists of 17 statements and is scored from 0 to 17 which higher scores indicating more positive levels of morale. No minimum clinically important difference (MCID) score is reported (McGlothlin & Lewis, 2014) which makes it impossible to evaluate any clinical impact or relevance. The average scores on this scale are well above the mid-point of nine and indicate that both groups expressed a high degree of morale. The difference is one point, and it is impossible to say that this represents a substantial, psychologically important difference.

For depression, the Geriatric Depression Scale was used (Yesavage and Sheikh, 1986). This scale consists of 15 questions and is scored from 0 to 15, with higher scores indicative of depression. The guidance for the scale provided by the developers states that a score of five or more “suggests” depression. Again, no MCID is reported. In the Cohen data, the singing group on average scored one and the control group two, so clearly neither group was depressed. It is impossible to say what a difference of one point on this scale means, but certainly, there is no indication that members of the comparison group were more depressed, or at a greater “risk for depression” (Cohen et al., 2006, p. 733).

For loneliness, the UCLA Loneliness Scale was used (Russell, 1996). A 20-item scale designed to measure one’s subjective feelings of loneliness as well as feelings of social isolation. Participants rate each item on a scale from 1 (Never) to 4 (Often) to give scores from 20 to 80, with higher scores indicating greater loneliness and a mid-point of 50. Scores between 20 and 34 are taken to mean a low degree of loneliness, 35–49 a moderate degree of loneliness, 50–64 a moderately high degree of loneliness, and 65–80 a high degree of loneliness. No MCID score is reported. In the Cohen et al. study, the means for both groups were well below the mid-point and on the boundary between low and moderate loneliness. The difference is only two points, and as with the measures of morale and depression, this difference is unlikely to be psychologically or clinically important.

In addition to these four issues identified through a critical reading of the Cohen et al. (2006) research, assessments were also made using the JBI critical appraisal tool for quasi-experimental studies. Of the nine items in this tool, three authors agreed on positive assessments for five questions (i.e. 1. Is it clear in the study what is the “cause” and what is the “effect”? 3. Were the participants included in any comparisons receiving similar treatment/care other than the exposure or intervention? 4. Was there a control group? 5. Were their multiple measurements of the outcome both pre and post the intervention/

exposure? 7. Were the outcomes of participants included in any comparisons measured in the same way?). However, negative assessments were agreed for the remaining four questions (i.e. 2. Were the participants included in any comparisons similar? 6. Was follow up complete and if not, were differences between groups in their follow-up adequately described and analyzed? 8. Were outcomes measured in a reliable way? Was appropriate statistical analysis used?). The negative picture emerging from the JBI tool is consistent with the four main critical issues presented earlier. In our view, the weaknesses of the Cohen et al. study (Cohen et al., 2006, 2007) mean that it should not have been included in any form of evidence review, and certainly not systematic reviews which involve careful appraisal of research quality. Nevertheless, the 2006 paper reporting on the year 1 follow-up results has been widely cited in subsequent evidence reviews.

### ***Citation of the Cohen et al. study in arts-focused evidence reviews***

The Cohen et al. paper has been cited 710 times according to Google Scholar (1 November 2023) and is included in no fewer than 32 evidence reviews of various kinds. Seventeen arts-focused reviews consider research on the benefits of participation in creative arts and music programmes and report details of the Cohen et al. (2006) work. A Scopus search (1 November 2023) identified 309 sources citing Cohen et al. and 22 evidence reviews including 12 art-focused evidence reviews. A further search using BASE (1 November 2023) identified 452 sources citing Cohen et al. and 24 evidence reviews including 12 arts-focused reviews. Details of the search results from Google Scholar, Scopus and BASE can be found in Supplementary Table 1. None of the searches identified (anonymised for peer review), which provided the starting point for this exercise.

Seven of the arts-focused reviews identified by Google Scholar, Scopus and BASE are not considered further due to the specific concerns of the reviews, the nature of the arts interventions considered or the population focus. For example, Zeilig et al. (2014) are concerned with arts interventions for people with dementia. Archibald and Kitson (2020) focus on “arts for awareness, communication and knowledge translation in older adulthood” and is not concerned with health outcomes. Galassi et al. (2022) is concerned with the role of creativity and art therapy for healthy aging. The Cohen et al. study does not meet the review inclusion criteria but is cited without commentary.

The remaining 15 “non-arts” reviews, identified by Google Scholar, Scopus and BASE are not considered here as their concerns extend beyond arts interventions or have a specific health or wellbeing focus. For example, the Cohen study is included in reviews of “social” or “community” programmes (e.g. Ghiga et al., 2020; Paquet et al., 2023), or “non-pharmacological” interventions (e.g. Sau-Fung et al., 2023). The Cohen-Mansfield and Perach (2015) review considers a wide range of interventions to address loneliness among older people, which is only one outcome of concern in Cohen et al. (2006).

Table 3 provides details of the treatment of the Cohen et al. (2006) findings in 11 evidence reviews, with a focus on creative arts, music and singing interventions for health. These reviews were published over 13 years from 2010 to 2023, and some are authored by active and widely cited researchers in the field of arts, music, and health (e.g. Daykin, Creech, and Särkämö).

The reviews included in Table 3 do vary in character and quality, and standards for the conduct and reporting of reviews have improved over the period of time covered by these

**Table 3.** Effects of singing reported by Cohen et al. (2006) and cited in 11 arts evidence reviews.

Effects Reviews	Overall health Yes	GP visits Yes	Medication Yes	Falls Yes	Other problems Yes	Morale Yes	Depression No	Loneliness Yes	Activities Yes
Bellazzecca 2022	Yes					Yes		Yes	Yes
Castora-Binkley 2010	Yes	Yes	Yes	Yes		Yes	No	Yes	
Clements-Cortez 2015	Yes	Yes	Yes	Yes		Yes		Yes	Yes
Anonymised 2010	Yes	Yes	Yes	Yes	Yes				
Creech 2023		Yes	Yes	Yes	Yes	Yes		Yes	Yes
Daykin 2018						Yes	Yes	Yes	
Gick 2011	Yes	Yes	Yes	Yes					
Lehmberg 2010	Yes	Yes	Yes			singers 'more positive responses on mental health measures'			Yes
Noice 2014	Yes	Yes	Yes	Yes					Yes
Särkämö 2018	Yes					Yes		Yes	Yes
Sheppard 2020	Yes	Yes		Yes			Yes		Yes

Green = correct report of finding, Orange = incorrect report of finding.

reviews (from 2010 to 2023). For example, PRISMA guidelines for the reporting of systematic reviews were first issued in 2009 (Page et al., 2021; Shamseer et al., 2015), and the online PROSPERO database for the prospective registration of systematic reviews was launched in 2011 (Page et al., 2018).

Three reviews are described as systematic, five as “reviews”, two as “critical reviews”, and one as a systematic mapping of research on singing and health. All three systematic reviews (Creech et al., 2023; Daykin et al., 2018; Sheppard & Broughton, 2020) include a PRISMA-style diagram. Only one of the three systematic reviews was, however, pre-registered in PROSPERO (Daykin et al., 2018).

In Table 3, the first column lists the 11 arts and music-focused evidence reviews that include the Cohen et al. (2006) study, in alphabetic order by first author. Along the top of this table are nine measures employed in the study (with the social activities variables combined) with “yes” indicating a reported difference by Cohen et al. in favour of the singing group. In rows for each of the reviews, green entries indicate an accurate mention in the review of differences reported by Cohen et al. The orange boxes indicate errors, which arise where reviews claim differences between the singing and comparison groups in depression at follow-up, which are not reported by Cohen et al (Daykin et al., 2018; Sheppard & Broughton, 2020).

All reviews, with the exception of (removed to anonymise) offer no critical commentary on the methods and findings of the Cohen et al. study and simply accept the results reported at face value. The systematic reviews do offer some evaluation of the Cohen et al. study, or comment on general methodological issues in reflecting on all of the studies included in the review. In Daykin et al. (2018), however, the quality assessment of the Cohen et al. study is positive, and in their more detailed report for the *What Works Wellbeing Centre* (Daykin et al., 2016) on which the 2018 peer-reviewed paper is based, they give an overall rating of “Good” for the Cohen et al. study, based on an assessment guided by GRADE criteria (see <https://www.gradeworkinggroup.org/>). This assessment is equivalent to that given to the Coulton et al. (2015) pragmatic randomised controlled trial

included in the systematic review, which is clearly more rigorous. Särkämö (2018) in his “critical review” is content to take the findings from Cohen et al. at face value, but in the general reflections in his paper he points to a need for improved research on music interventions for health.

## Discussion

In the present paper, we have taken one research study on singing and health (Cohen et al., 2006) and examined the way in which it has been treated in 11 subsequent reviews. We have demonstrated that, in general, the authors of these reviews appear to be content to take the findings from the Cohen et al. study at face value without subjecting them to rigorous scrutiny.

Of course, this raises questions about the treatment of all the other research studies included in these reviews. We acknowledge that it would be a very time-consuming task to check every original paper and that, ideally, original papers that have undergone peer-review should be considered trustworthy. However, since our present paper demonstrates that the latter is not the case, it raises questions about the specific and general conclusions drawn in the reviews and any implications identified for policy and practice.

### *Why so many reviews?*

In the present paper, we have considered 11 reviews which have included Cohen et al. (2006) and which focus on research on creative arts and music. However, as we noted earlier, a further 13 reviews focused on a broader range of social or community programmes also have included the Cohen et al. study (Cohen et al., 2006). There are some obvious answers as to why there should be so many reviews. The reviews cover a period of 13 years from 2010 to 2023, and so new research may have appeared to warrant updated reviews. Furthermore, each review is concerned with a different specific research question, and, thus, may cover a different body of literature, and they vary in their search strategies. Such factors also account for the fact that recent reviews that might be expected to include the Cohen et al. (2006) paper, do not include it. Reagon et al. (2016), for example, in a systematic review of singing and health, focus specially on controlled studies that include a validated outcome measure of health-related quality of life, which is not the case for Cohen et al. (2006). Campbell et al. (2022) report a narrative review of group singing and physical and psychological wellbeing, and Yi and Kim (2023) report a scoping review of community music activities for wellbeing. Both fail to identify the Cohen et al. (2006) study, most likely due to their search strategy, which not surprisingly focuses on finding studies on singing. For the Cohen et al. (2006) paper, “singing” does not appear in the paper’s title or abstract, nor is it a keyword. McQuade and O’Sullivan (2023) also report a systematic review of arts and creativity in later life, which might be expected to include the Cohen et al. (2006) study, but in this case, they include studies appearing from 2013 onwards.

A further factor worth noting is that several recent reviews appear to have been conducted without awareness of previous reviews. Särkämö (2018) and Daykin et al. (2018), for example fails to cite any of the previous reviews considered here. Creech et al. (2023) and Sheppard and Broughton (2020) do not cite Särkämö (2018) or Daykin

et al. (2018). The picture is better, however, for Bellazzecca et al. (2022) who cite Noice et al. (2014) and Daykin et al. (2018), but not Särkämö (2018), Creech et al. (2020) or Sheppard and Broughton (2020).

Two systematic reviews considered were not pre-registered (Creech et al. 2020; Sheppard & Broughton, 2020), despite the availability of PROSPERO for prospective registration. Thus, teams working on the same topic concurrently would have no knowledge of related reviews-in-progress.

### ***Why are reviews so uncritical of the Cohen et al. study?***

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 (removed to anonymise). The question can be applied to all the reviews we consider here with the exception of (removed to anonymise). In the non-systematic reviews, it is difficult not to draw the conclusion that the authors were lax in summarising the findings of the study or simply took on trust the conclusions reached by Cohen et al. (2006).

Quality screening or consideration of sources of bias is generally lacking or limited in most of the reviews. Sheppard and Broughton (2020), for example, are content to state that:

For inclusion in this review, each selected study had to have been subject to peer review processes prior to publication, and had to present a clear, consistent methodology, which were both taken as indicators of research quality. (p.3)

Daykin et al. (2018) are more careful in their evaluation of the studies guided by quality scales (i.e. GRADE for quantitative studies and CERQual for qualitative research) but they give an inflated assessment of the Cohen et al. (2006) study as being of “high” quality, most likely because it includes a comparison group. The Coulton et al. (2015) pragmatic randomised controlled trial is also rated as “high” but as an RCT is clearly of greater quality than the Cohen et al. (2006) research.

None of the reviews we consider offers a more radical critique and raise questions regarding the limitations and weaknesses we identify here:

- obvious sources of bias in the study due to lack of blinding and the details of the purpose of the study given on recruitment,
- failure to address the potential bias introduced by attrition from baseline to one-year follow-up,
- inadequacies of the statistical strategy adopted, and, finally,
- lack of attention to the minor substantive changes reported in the self-report measures of overall health and mental wellbeing.

Both Hammersley (2020) and Ioannidis (2016) offer possible explanations of why reviews can be uncritical. Hammersley (2020) points out that systematic reviews involve a number of stages, and the whole process can be very time-consuming. It may be that if considerable time is devoted to ensuring that the search and selection process is systematic and comprehensive, then less time is available for rigorous

evaluation beyond the application of quality checklists. Once the stages of searching and selecting studies and quality screening are completed, reviewers may feel that their work is done and that all that remains is to summarise the studies and provide a narrative synthesis.

Ioannidis (2016) also raises concerns about the possible role of “vested interests” of academics who conduct systematic reviews and meta-analyses. In his view:

Ideally, people who have no stake in the results should perform systematic reviews and meta-analyses, excluding not only those with financial conflicts of interest but even those who are content experts in the field. According to this line of argument, content experts can and should be consulted, but they should not be authors. (p. 495)

In the field of arts and health, the challenge raised by Ioannidis is whether the starting point of a review team is one of “dispassionate enquiry and scepticism” or whether team members already believe that the arts have benefits for health and wellbeing? If the former, a review team may adopt a more critical approach to the research studies included. If the latter, however, a review may be conducted with the intention of showcasing positive evidence.

A further concern of review authors may be to advocate for supportive policy development, further funding for research, and the practical implementation and wider scaling up of arts for health programmes. This issue is a potential source of bias in several of the reviews considered here, where reviewers include primary studies which they themselves have conducted (e.g. Daykin et al., 2018, Särkämö, 2018, and Creech et al. (2020)).

### **Limitations**

There are limitations to the work we report here. We have only undertaken an analysis of one target paper by Cohen et al. (2006) and considered the way it is treated in 11 creative arts and music focused evidence reviews. However, three earlier papers (removed to anonymise) have revealed the same concerns regarding the lack of critical scrutiny of primary research on arts therapies and arts interventions in subsequent evidence reviews.

A further possible limitation is that only three search engines, Google Scholar, Scopus and BASE were used to identify citations of the Cohen et al. study in subsequent evidence reviews. Recent information science investigations of Google Scholar as a multi-disciplinary search engine have demonstrated its superior performance over alternative academic bibliographic databases providing a citation function (Gussenbauer, 2019; Martínmartín et al., 2021). However, Gussenbauer and Haddaway (2020) suggest that in conducting searches for systematic reviews, that Google Scholar is supplemented by further search engines and specifically recommends BASE. We found that Google Scholar identified more subsequent sources and reviews citing Cohen et al. than SCOPUS and BASE and did not locate further sources in addition to those already revealed by Google Scholar. We are therefore confident that our search strategy did not miss any evidence reviews in the literature citing Cohen et al. (2006).

## Conclusions

The present paper has demonstrated several issues regarding the Cohen et al. (2006) paper as one example of a paper that has been included in creative arts and music focused evidence reviews.

Specifically, we have identified concerns regarding methodology, findings, and conclusions in the original Cohen et al. paper (Cohen et al., 2006) and a subsequent lack of critical scrutiny in 11 evidence reviews, leading to inaccurate conclusions.

Thus, the findings in the present paper concur with the findings in our three earlier papers (Clift, et al., 2022; Grebosz-Haring, et al., 2022; Grebosz-Haring, et al., 2023) in which we revealed the same concerns regarding the lack of critical scrutiny of primary research on arts therapies and arts activities in evidence reviews.

Our work therefore leads to the following recommendations:

- The Cohen et al. paper (2006) should not be considered suitable for inclusion in future evidence reviews or in background sections of original research papers – without thorough critical perspectives.
- Further review studies following the innovative method demonstrated in this and previous papers are needed to assess the accuracy and credibility of evidence reviews in the field of arts and health.
- Systematic reviews should be planned following the current guidelines (Aromataris & Munn, 2020, Higgins et al., 2023); protocols should be pre-registered in PROSPERO and reviews reported following current PRISMA guidelines. Increasingly, there is an expectation that scoping review protocols should be published, and an extension of PRISMA guidelines for reporting scoping reviews should be followed.
- Peer review of reports of evidence reviews needs to be rigorous and involves 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 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. For example, it is a matter of serious concern that the study conducted by Cohen et al. (2006) has never been replicated and improved upon in the form of an RCT.
- Evidence reviews 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 of all forms with an appropriate degree of caution.
- The field of arts for health research is still at a preliminary stage and, therefore, policy development, practical implementation, and wider scaling up of arts for health programmes should be considered with reasonable caution.
- Further original research of high quality is needed to investigate the effects and benefits of arts for health interventions and to build a solid body-of-research suitable to include in high-quality reviews and meta-analyses.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

The author(s) reported that there is no funding associated with the work featured in this article.

## References

- Archibald, M., & Kitson, A. L. (2020). Using the arts for awareness, communication and knowledge translation in older adulthood: A scoping review. *Arts & Health, 12*(2), 99–115. <https://doi.org/10.1080/17533015.2019.1608567>
- Aromataris, E., & Munn, Z. (Eds). (2020). *JBI manual for evidence synthesis*. Joanna Briggs Institute. <https://doi.org/10.46658/JBIMES-20-01>
- Bailey, B. A., & Davidson, J. W. (2002). Adaptive characteristics of group singing: Perceptions from members of a choir for homeless men. *Musicae Scientiae, 6*(2), 221–256. <https://doi.org/10.1177/102986490200600206>
- Beck, R. J., Cesari, T. C., Yousefi, A., & Enamoto, H. (2000). Choral singing, performance perception, and immune system changes in salivary immunoglobulin a and cortisol. *Music Perception, 18*(1), 87–106. <https://doi.org/10.2307/40285902>
- Bellazzecca, E., Teasdale, S., Biosca, O., & Skelton, D. A. (2022). The health impacts of place-based creative programmes on older adults' health: A critical realist review. *Health & Place, 76*, 102839. <https://doi.org/10.1016/j.healthplace.2022.102839>
- Bone, J. K., & Fancourt, D. (2022). *Arts, culture & the brain: A literature review and new epidemiological analyses*. Arts Council England. <https://www.artscouncil.org.uk/arts-culture-brain>
- Campbell, Q., Bodkin-Allen, S., & Swain, N. (2022). Group singing improves both physical and psychological wellbeing in people with and without chronic health conditions: A narrative review. *Journal of Health Psychology, 27*(8), 1897–1912. <https://doi.org/10.1177/13591053211012778>
- Castora-Binkley, M., Noelker, L., Prohaska, T., & Satariano, W. (2010). Impact of arts participation on health outcomes for older adults. *Journal of Aging, Humanities, and the Arts, 4*(4), 352–367. <https://doi.org/10.1080/19325614.2010.533396>
- Clements-Cortés, A. (2015). Clinical effects of choral singing for older adults. *Music & Medicine, 7*(4), 7–12. <https://doi.org/10.47513/mmd.v7i4.437>
- Clift, S. (2020). The need for robust appraisal of research in arts and health: A case of the Emperor's new clothes. In Pritchard, S. (Ed.), *The need for robust appraisal of research in arts and health: A case of the Emperor's new clothes*. <https://colouringinculture.org/blog/the-need-for-robust-appraisal-of-research-in-arts-and-health-4/>
- Clift, S., Gilbert, R. & Vella-Burrows, T. (2016). *A choir in every care home: A review of research on the value of singing for older people*. Live Music Now!. <https://achoirineverycarehome.files.wordpress.com/2016/04/wp6-research-review-v2-1.pdf>
- 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. The need for robust critique of arts and health research: An examination of the Goldbeck and Ellerkamp. *Approaches: An Interdisciplinary Journal of Music Therapy*, published 11 August, <https://approaches.gr/clift-a20220811/>
- Clift, S. M., & Hancox, G. (2001). The perceived benefits of singing: Findings from preliminary surveys of a university college choral society. *Journal of the Royal Society for the Promotion of Health, 121* (4), 248–256. <https://doi.org/10.1177/146642400112100409>



- Clift, S., Hancox, G., Staricoff, R. & Whitmore, C. (2008). *Singing and health: A systematic mapping and review of non-clinical research*. Canterbury Christ Church University. <https://www.artshealthresources.org.uk/docs/singing-health-a-systematic-mapping-review-of-non-clinical-research/>
- Clift, S., Nicol, J., Raisbeck, M., Whitmore, C. & Morrison, I. (2010). Group singing, wellbeing and health: A systematic mapping of research evidence. *UNESCO Observatory E-Journal*, 2(1), 1–25. <https://www.unescoejournal.com/wp-content/uploads/2020/03/2-1-11-clift-paper.pdf>
- 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://doi.org/10.1080/09548963.2021.1910492>
- Cohen-Mansfield, J., & Perach, R. (2015). Interventions for alleviating loneliness among older persons: A critical review. *American Journal of Health Promotion*, 29(3), 109–125. <https://doi.org/10.4278/ajhp.130418-LIT-182>
- Cohen, G., Perlstein, S., Chapline, J., Kelly, J., Firth, K. M., & Simmens, S. (2006). The impact of professionally conducted cultural programs on the physical health, mental health, and social functioning of older adults. *The Gerontologist*, 46(6), 726–734. <https://doi.org/10.1093/geront/46.6.726>
- Cohen, G., Perlstein, S., Chapline, J., Kelly, J., Firth, K. M., & Simmens, S. (2007). The impact of professionally conducted cultural programs on the physical health, mental health, and social functioning of older adults. *Journal of Aging, Humanities, and the Arts*, 1(1–2), 5–22. <https://doi.org/10.1080/19325610701410791>
- Coulton, S., Clift, S., Skingley, A., & Rodriguez, J. (2015). Effectiveness and cost-effectiveness of community singing on mental health-related quality of life of older people: Randomised controlled trial. *British Journal of Psychiatry*, 207(3), 250–255. <https://doi.org/10.1192/bjp.bp.113.129908>
- Creech, A., Larouche, K., Generale, M., & Fortier, D. (2023). Creativity, music, and quality of later life: A systematic review. *Psychology of Music*, 51(4), 1080–1100. <https://doi.org/10.1177/0305735620948114>
- Cutler, D. (2019). *Around the world in 80 creative ageing projects*. Baring Foundation. <https://baringfoundation.org.uk/resource/around-the-world-in-80-creative-ageing-projects/>
- Daykin, N., Julier, G., Tomlinson, A., Meads, C., Mansfield, L., Payne, A., Duffy, L. G., Lane, J., D’Innocenzo, G., Burnett, A., Kay, T., Dolan, P., Testoni, S., & Victor, C. (2016). *Music, singing and wellbeing in healthy adults: A systematic review*. What Works Wellbeing. <https://whatworkswellbeing.org/wp-content/uploads/2020/01/1-systematic-review-healthy-adult-music-singing-wellbeing-nov2016final.pdf>
- Daykin, N., Mansfield, L., Meads, C., Julier, G., Tomlinson, A., Payne, A., Duffy, L. G., Lane, J., D’Innocenzo, G., Burnett, A., Kay, T., Dolan, P., Testoni, S., & Victor, C. (2018). What works for wellbeing? A systematic review of wellbeing outcomes for music and singing in adults. *Perspectives in Public Health*, 138(1), 39–46. <https://doi.org/10.1177/1757913917740391>
- Ellis, P. D. (2010). *The essential Guide to effect sizes*. Cambridge University Press.
- 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://apps.who.int/iris/handle/10665/329834>
- Fancourt, D., Warren, K., & Aughterson, H. (2020). *Evidence summary for policy: The role of arts in improving health & wellbeing*. University College London. <https://www.gov.uk/government/publications/evidence-summary-for-policy-the-role-of-arts-in-improving-health-and-wellbeing>
- Galassi, F., Merizzi, A., D’Amen, B., & Santini, S. (2022). Creativity and art therapies to promote healthy aging: A scoping review. *Frontiers in Psychology*, 13, 906191. <https://doi.org/10.3389/fpsyg.2022.906191>
- Ghiga, I., Pitchforth, E., Lepetit, L., Miani, C., Ali, G.-C., & Meads, C. (2020). The effectiveness of community-based social innovations for healthy ageing in middle- and high-income countries: A systematic review. *Journal of Health Services Research & Policy*, 25(3), 202–210. <https://doi.org/10.1177/1355819619888244>
- Gick, M. L. (2011). Singing, health and well-being: A health psychologist’s review. *Psychomusicology: Music, Mind, and Brain*, 21(1–2), 176–207. <https://doi.org/10.1037/h0094011>

- Gorard, S. (2021). *How to make sense of statistics*. Sage.
- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Grebosz-Haring, K., Thun-Hohenstein, L., Schuchter-Wiegand, A. K., Bathke, A. & Clift, S. (2023). The need for robust critique of arts and health research: Dance movement therapy, adolescent girls, and depression. *Annals of the New York Academy of Sciences, Online Publication* 25 May, <https://doi.org/10.1111/nyas.15006>
- 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>
- Greenhalgh, T., Thorne, S., & Malter, 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>
- Gussenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177–184. <https://doi.org/10.1007/s11192-018-2958-5>
- Gussenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2), 181–217. <https://doi.org/10.1002/jrsm.1378>
- Hallam, S., & Himonides, E. (2022). *The power of music: An exploration of the evidence*. Open Book. <https://books.openbookpublishers.com/10.11647/obp.0292.pdf>
- Hammersley, M. (2020). Reflections on the methodological approach of systematic reviews. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic reviews in educational research* (pp. 23–39). Springer VS.
- Higgins, J., James, T., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2023). *Cochrane Handbook for Systematic Reviews of Interventions (Version 6.4)*. <https://training.cochrane.org/handbook/current>
- 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://doi.org/10.1111/1468-0009.12210>
- Johnson, J., Stewart, A. L., Acree, M., Nápoles, A. M., Flatt, J. D., Max, W. B., & Gregorich, S. E. (2020). A community choir intervention to promote wellbeing among diverse older adults: Results from the community of voices trial. *The Journals of Gerontology: Series B*, 75(3), 549–559. <https://doi.org/10.1093/geronb/gby132>
- Kreutz, G., Bongard, S., Rohrmann, S., Grebe, D., Bastian, H. G., & Hodapp, V. (2004). Effects of choir singing or listening on secretory immunoglobulin A, cortisol and emotional state. *Journal of Behavioral Medicine*, 27(6), 623–635. <https://doi.org/10.1007/s10865-004-0006-9>
- Lawton, M. P. (1975). The Philadelphia Geriatric Center morale scale: A revision. *Journal of Gerontology*, 30(1), 85–89. <https://doi.org/10.1093/geronj/30.1.85>
- Lehmberg, L. J., & Fung, V. (2010). Benefits of music participation for senior citizens: A review of the literature. *Music Education Research International*, 4, 19–30. <http://cmer.arts.usf.edu/content/articfiles/3122-MERI04pp.19-30.pdf>
- 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>
- Martínmartín, A., Thelwall, M., Orduna-Malea, E., & Lópezcózar, E. D. (2021). Google Scholar, Microsoft academic, Scopus, dimensions, web of science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871–906. <https://doi.org/10.1007/s11192-020-03690-4>
- Maury, S., & Rickard, N. (2022). The benefits of participation in a choir and an exercise group on older adults' wellbeing in a naturalistic setting. *Musicae Scientiae*, 26(1), 144–171. <https://doi.org/10.1177/1029864920932633>

- McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *JAMA*, 312(13), 1342–1343. <https://doi.org/10.1001/jama.2014.13128>
- McQuade, L., & O'Sullivan, R. (2023). Examining arts and creativity in later life and its impact on older people's health and wellbeing: A systematic review of the evidence. *Perspectives in Public Health*. Online first. <https://doi.org/10.1177/17579139231157533>.
- 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>
- Moore, A., Fisher, E., & Eccleston, C. (2023). Flawed, futile and fabricated – features that limit confidence in clinical research in pain and anaesthesia: A narrative review. *British Journal of Anaesthesia*, 130(3), 287–295. <https://doi.org/10.1016/j.bja.2022.09.030>
- Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology*, 18(5), 1–9. <https://doi.org/10.1186/s12874-017-0468-4>
- Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence-Based Practice*, 21(4), 125–27. <https://doi.org/10.1136/ebmed-2016-110401>
- Nicols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, 135(2), 151–166. <https://doi.org/10.3200/GENP.135.2.151-166>
- Noice, T., Noice, H., & Kramer, A. F. (2014). Participatory arts for older adults: A review of benefits and challenges. *The Gerontologist*, 54(5), 741–753. <https://doi.org/10.1093/geront/gnt138>
- Page, M. J., McKenzie, J. E., Bossuyt, M., Boutron, I., Hoffman, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grinshaw, J. M., Hróbjartsson, A., Lalu, M. M., Tianjing, L., Loder, E. W., Mayo-Wilson, E., & McDonald, S. & Moher, D., Whiting, P. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The British Medical Journal*, 372, n71. <https://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(1), 32. <https://doi.org/10.1186/s13643-018-0699-4>
- Paquet, C., Whitehead, J., Shah, R., Adams, A. M., Dooley, D., Spreng, R. N., Aunio, A.-L., & Dubé, L. (2023). Social prescription interventions addressing social isolation and loneliness in older adults: Meta-review integrating on-the-ground resources. *Journal of Medical Internet Research*, 25, e40213. <https://doi.org/10.2196/40213>
- Reagon, C., Galea, N., Enright, S., Mann, M., & van Deursena, R. (2016). A mixed-method systematic review to investigate the effect of group singing on health related quality of life. *Complementary Therapies in Medicine*, 27, 1–11. <https://doi.org/10.1016/j.ctim.2016.03.017>
- Russell, D. (1996). The UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment*, 66(1), 20–40. [https://doi.org/10.1207/s15327752jpa6601\\_2](https://doi.org/10.1207/s15327752jpa6601_2)
- Särkämö, T. (2018). Cognitive, emotional, and neural benefits of musical leisure activities in aging and neurological rehabilitation: A critical review. *Annals of Physical and Rehabilitation Medicine*, 61(6), 414–418. <https://doi.org/10.1016/j.rehab.2017.03.006>
- Sau-Fung, Y. D., Wai-Chi, L. P., Sin-Yi, L. R., Frank, K., Alice, C., & Pll, W. W. (2023). Effects of non-pharmacological interventions on loneliness among community-dwelling older adults: A systematic review, network meta-analysis, and meta-regression. *International Journal of Nursing Studies*, S0020-7489(23)00089–5. Advance online publication. <https://doi.org/10.1016/j.ijnurstu.2023.104524>
- 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. *The British Medical Journal*, 349(jan02 1), g7647. <https://doi.org/10.1136/bmj.g7647>
- Sheppard, A., & Broughton, M. C. (2020). Promoting wellbeing and health through active participation in music and dance: A systematic review. *International Journal of Qualitative Studies on Health and Well-Being*, 15(1), 1732526. <https://doi.org/10.1080/17482631.2020.1732526>

- 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>
- Yesavage, J. A., & Sheikh, J. I. (1986). Geriatric depression scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health*, 5(1–2), 165–173. [https://doi.org/10.1300/J018v05n01\\_09](https://doi.org/10.1300/J018v05n01_09)
- Yi, S. Y., & Kim, A. J. (2023). Implementation and strategies of community music activities for well-being: A scoping review of the literature. *International Journal of Environmental Research and Public Health*, 20(3), 2606. <https://doi.org/10.3390/ijerph20032606>
- Zeilig, H., Killick, J., & Fox, C. (2014). The participative arts for people living with a dementia: A critical review. *International Journal of Ageing & Later Life*, 9(1), 7–34. <https://doi.org/10.3384/ijal.1652-8670.14238>