Université de Montréal

Efficacy of Musical Interventions on Social, Maladaptive and Language Outcomes in Children with Autism Spectrum Disorder

A systematic review and meta-analysis

Par Nathalie Roth

Département de psychologie
Faculté des arts et sciences

Mémoire présenté en vue de l'obtention du grade de Maîtrise
en Psychologie
option recherche

Juillet, 2019

Université de Montréal
Département de psychologie, faculté des arts et sciences

Ce mémoire intitulé

Efficacy of Musical Interventions on Social, Maladaptive and Language Outcomes in Children with Autism Spectrum Disorder

A systematic review and meta-analysis

Présenté par

Nathalie Roth

A été évalué par un jury composé des personnes suivantes

Krisa Hyde

Directrice de recherche

Simona Brambati

Codirectrice

Nathalie Gosselin

Membre du jury

Arnaud Saj

Membre du jury

Résumé

Un nombre croissant d'études indique que les interventions musicales (IMs) ont des effets bénéfiques sur les enfants et adolescents atteints d'un trouble du spectre de l'autisme (TSA). À ce jour, aucune revue systématique utilisant une approche méta-analytique n'a investigué l'efficacité des IMs sur trois des symptômes prédominants du TSA, soit le fonctionnement social, les comportements mal-adaptés et le langage. Dans ce mémoire, tous les 17 articles scientifiques comparant les IMs à des interventions non-musicales (INMs) ont été décrits systématiquement et évalués selon les lignes directrices de CONSORT. Onze études répondaient aux critères d'inclusion des présentes méta-analyses, chacune rapportant des mesures d'évaluations longitudinales ou d'évaluations pendant les sessions. Les résultats des méta-analyses ont démontré un bénéfice des IMs chez le TSA, particulièrement pour les mesures des comportements sociaux mal-adaptés. Une comparaison entre les IMs et les INMs indiquait un avantage pour les IMs comparément aux INMs en ce qui concerne les comportements sociaux, mais aucun avantage pour les comportements mal-adaptés non-sociaux et le langage. La revue systématique a révélé des lacunes méthodologiques des études évaluées, telles que des tailles d'échantillons restreintes, des durées et intensités d'interventions limitées, un manque d'information à propos des échantillons et des critères d'appariement, puis le biais d'attrition. La combinaison de cette revue systématique et des méta-analyses a permis une mise à jour de l'évaluation des preuves de l'efficacité des IMs pour les jeunes ayant un TSA, ainsi que de donner des recommandations aux chercheurs et cliniciens afin d'améliorer la pratique dans ce domaine.

Mots-clés : autisme, musique, interventions musicales, fonctionnement social, comportements mal-adaptés, langage, méta-analyse, revue systématique

Abstract

There is considerable interest in using music interventions (MIs) to address core impairments present in children and adolescents with autism spectrum disorder (ASD). An increasing number of studies suggest that MIs have positive outcomes in this population, but no systematic review employing meta-analysis to date has investigated the efficacy of MIs across three of the predominant symptoms in ASD, specifically social functioning, maladaptive behaviors and language impairments. In this thesis, all available peer-reviewed studies (*n* = 17) comparing active MIs to non-music interventions (NMIs) were systematically summarized and evaluated using quality assessment based on the CONSORT statement. Eleven studies fulfilled inclusion criteria for meta-analysis, which differentiated between within-session and longitudinal outcomes. The quantitative analyses results supported the effectiveness of MI in ASD, and particularly for measures sensitive to social maladaptive behaviors. Comparing between MI and NMI, the results were generally suggestive of benefits of MI over NMI for social outcomes, but did not find benefits for non-social maladaptive behaviors or language outcomes. The systematic review revealed important methodological issues present in these studies, such as small sample sizes, restricted durations and intensities of interventions, missing sample information and matching criteria, and attrition bias. Together, the combined systematic review and meta-analyses provided an up-to-date evaluation of the evidence for MI's benefits in ASD children and provide key recommendations for future clinical interventions and research about best practice in the domain.

Keywords : Autism, music, music intervention, meta-analysis, systematic review, social functioning, maladaptive behavior, language

Table des matières

List of Tables

List of Figures

List of Abbreviations

ABA VB: Applied Behavior Analysis Verbal Behavior

ADI-R: Autism Diagnostic Interview-Revised

ADOS: Autism Diagnostic Observation Schedule

ASD: Autism spectrum disorder

ASSP: Autism Social Skills Profile

AVB: Analysis verbal behavior

CARS: Childhood Autism Rating Scale

CARS-BR: Childhood Autism Rating Scale adapted for use in Brazil

CCC-2: Children's Communication Checklist-2

DSB: Detrimental Social Behaviors

DSLM: Developmental speech and language through music

ESCS: Early Social Communication Scale

IQ: Intelligence quotient

JTAT: Joint Attention Test

MBCDI-WG: MacArthur Bates Communicative Development Inventory, Words and Gestures

MBCT: Melodic based communication training

MI: Music intervention

MT: Music therapy

NMI: Non-music intervention

PDDBI: Pervasive Developmental Disorder Behavior Inventory

PPC: Pre-post-control

PPVT: Peabody Picture Vocabulary Test

RCT: Randomized controlled trial

SCQ: Social Communication Questionnaire

SES: Socio-economic status

SMD: Standardized mean difference

SP: Social Participation

SR: Social Reciprocity

SRS: Social Responsiveness Scale

SSRS: Social Skills Rating Scale

TD: Typically developing

VABS: Vineland Adaptive Behaviour Scale

VABS-MB: Vineland Adaptive Behavior Scale – Maladaptive Behaviours

VPES: Verbal Production Evaluation Scale

VSEEC: Vineland Social-Emotional Early Childhood Scales

Acknowledgements

I would like to express my gratitude to my supervisor Krista Hyde, who allowed me to contribute to the excellent research in her lab, providing me with the unique opportunity to write this Master thesis at the Université de Montréal. I would also like to thank my co-supervisor, Simona Brambati, for her kind and heartfelt support during the writing process.

I want to express a special thank you to Nick Foster, who provided me with excellent and relentless mentoring throughout the entire research and writing process that made this study and thesis possible.

I would also like to thank Kevin Jamey for his kind contribution to this paper through careful assessment and sharing his own experiences.

A special thanks also goes out to my colleagues at the ABCD Lab, Alexa Meilleur and Sarah-Maude Coll, and to the extended BRAMS Lab, for their kindness, humor and sympathy during this shared research time.

Finally, I want to thank my family and friends, both near and far, for believing in me and showing patience and understanding throughout this process.

Introduction

General Introduction

A growing number of studies report that music interventions (MIs) have positive outcomes in clinical populations such as children and adolescents with autism spectrum disorder (ASD). Effectively, 12% of interventions in ASD employ music (Bhat & Srinivasan, 2013). With the increasing popularity of MIs, there is a growing need for rigorous assessment of their efficacy for symptom treatment. To our knowledge, no systematic review employing meta-analysis to date has investigated the efficacy of MIs across three of the predominant symptoms in ASD children and adolescents, specifically social functioning, maladaptive behaviors and language. Taking a comprehensive perspective on music interventions, the present study aims to inform practitioners on one hand on which strategies to prioritize in their work with this population, and researchers on another as to what elements of MI may achieve favorable outcomes in ASD children and adolescents. Accordingly, through systematic review and meta-analyses, all currently available research in ASD children and adolescents that compares MIs to non-musical interventions (NMI) is assessed here in terms of core social, behavioral and communicative outcomes.

Music Interventions

MI uses music to achieve therapeutic goals, such as improving social function and reducing problematic behaviors (Cogo-Moreira, de Avila, Ploubidis, & de Jesus Mari, 2013). MI and music therapy (MT) are frequently used synonymously (e.g. Geretsegger, Elefant, Mössler, & Gold, 2014; Jackson, 2003). However, MI can be implemented by any professional without any specific MT training (Maloy & Peterson, 2014), while MT is defined as the “clinical and evidence-based use of music interventions to accomplish individualized goals within a therapeutic relationship by a credentialed professional who has completed an approved music therapy program” (American Music Therapy Association, 2013). Therefore, MT is also a MI, but not every MI qualifies as MT. The scope of the present study, in an effort to be as comprehensive as possible, will be inclusive of all types of MIs studied in ASD.

The use of music in MIs can take the form of a background or contingent stimulus, or as an interactive activity that engages patients to use their voice, movement and music instruments in a way that meets their therapeutic needs (Kim, Wigram, & Gold, 2009). Music is believed to have beneficial effects in both neurotypical and clinical populations due to its multisensory characteristics, its communicational aspects and its enjoyability (Suthers & Niland, 2007). For example, in neurotypical children, some sources have reported that music has prosocial benefits (e.g. Kirschner & Tomasello, 2010; Miendlarzewska & Trost, 2014; Särkämö, Tervaniemi, & Huotilainen, 2013). A systematic review by Dumont, Syurina, Feron, and van Hooren (2017) assessed the effects of MI in social, communication, language, cognitive, motor and academic domains in school-aged children, but while their review is suggestive of beneficial effects of MIs on development, clear conclusions could not be drawn because of methodological heterogeneity and poor quality of the analyzed studies. In particular, the authors noted that outcomes were limited by small sample sizes, incoherent or missing control conditions and lack of adherence to randomized controlled trial (RCT) design.

MIs have been used effectively in various clinical populations such as dementia, schizophrenia and terminal illness (Tang & Vezeau, 2010). A summary of systematic reviews by Kamioka et al. (2014) showed that MIs are particularly effective in increasing social and global functioning in schizophrenia, gait and posture in Parkinson's disease and depressive symptoms, and sleep quality in major depressive disorder. Raglio et al. (2008) demonstrated that active MT alleviated behavioral and psychological symptoms in patients with moderate to severe dementia. He et al. (2018) found that listening to classical music decreased psychiatric symptoms in a schizophrenic sample. There is some evidence that MIs are also effective in ASD, a developmental disorder with rising prevalence (Accordino, Comer, & Heller, 2007; Autism Speaks, 2018; Geretsegger et al., 2014). Especially taking into consideration that ASD children are said to enjoy music (Gebauer, Skewes, Westphael, Heaton, & Vuust, 2014), it is not surprising that a rising number of interventions for this population employ music (Bhat & Srinivasan, 2013).

Key Outcomes in Autism Spectrum Disorder

ASD is a pervasive condition that first manifests in early childhood. Diagnosis is more common in boys than girls, and overall prevalence is estimated at about one in 59 children (Autism Speaks, 2018). Primary deficits in ASD include moderate to severe impairments in social-emotional functioning and communication and language skills. ASD is also characterized by repetitive, restricted and stereotyped behaviors, interests and activities. These deficits are present at an early age, cause significant impairment, and are not explained by other impairments or delays (American Psychiatric Association, 2013).

The social and emotional understanding of ASD children tends to be limited in varying degrees, impacting their ability to reciprocate appropriately (e.g. eye gaze, turn taking) (Katagiri, 2009). While not a diagnostic criterion in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013), individuals on the spectrum show different levels of language skills and some never develop verbal language (Dimitriadis & Smeijsters, 2011). When they learn to communicate, many manifest inappropriate use of words and gestures, and produce vocal sounds without context or communicative intent (Gattino, Riesgo, Longo, Leite, & Faccini, 2011). They develop obsessional and rigid behaviors, which they also require from other people (Geretsegger et al., 2014; Tsermentseli, Tabares, & Kouklari, 2018). These deficits are thought to greatly impact social functioning in ASD (Kouklari, Tsermentseli, & Auyeung, 2018), and can result in atypical, repetitive and sometimes self-injurious behavior (LaGasse, 2017). Other common maladaptive behaviors in ASD include temper tantrums, physical aggression, general irritability and disobedience (Samson, Hardan, Lee, Phillips, & Gross, 2015). There appears to be a strong relationship between maladaptive behaviors and communication skills in children with ASD (Fulton, Eapen, Črnčec, Walter, & Rogers, 2014), which makes it relevant to develop interventions that target both types of impairment.

MI in ASD

Several systematic reviews and meta-analyses (e.g. Geretsegger et al., 2014; Srinivasan, Eigsti, Neelly, & Bhat, 2016a; Whipple, 2004) recommend MI as a useful therapeutic tool to address deficits in ASD. Furthermore, some studies found that ASD children have enhanced auditory pitch perception and have stronger neural activations when exposed to songs compared to typically developing (TD) peers (LaGasse, 2014). Musical structure can bring a predictability that fulfills the need for anticipation in ASD, while allowing enough openness to address deficits such as joint attention, eye contact and turn-taking (Geretsegger et al., 2014; Geretsegger et al., 2015). Additionally, it has been noted that people with autism generally enjoy music (Gebauer et al., 2014) and even use it to manage their moods (Allen, Hill, & Heaton, 2009).

As explained by Kim et al. (2009), music represents a medium that facilitates the establishment of a meaningful relationship between interventionist and client. A so-called “musical attunement” is said to help create affective and reciprocal communication, which is often lacking in ASD children. Musical attunement means that the interventionist sensitively adapts the music activity to the client’s momentary needs in order to achieve a therapeutic goal (Kim, Wigram, & Gold, 2008). In a seminal systematic review on the effects of MIs on social behavior in ASD, Geretsegger et al. (2014) stated that engaging with music requires joint attention, eye contact and turn taking, which are social behaviors with which ASD children frequently struggle. The link between the intact music abilities in ASD, this population’s affinity for music, and the social behaviors that can be promoted through music makes MIs promising tools for enhancing social behaviors in ASD. Indeed, Geretsegger et al. (2014) found a moderate to high effect for MIs on social and communicative skills in ASD. However, other studies have found no significant effect for MI on social behavior and skills (e.g. Bieleninik et al., 2017; Sharda et al., 2018; Srinivasan, Eigsti, Gifford, & Bhat, 2016a; Srinivasan et al., 2016b).

Music has also been used to reduce challenging behavior in ASD, including self-injurious, aggressive and stereotypical behaviors (Bhat & Srinivasan, 2013; Devlin, Healy, Leader, & Reed, 2008; Orr, Myles, & Carlson, 1998; Pasiali, 2004). One way in which MI is proposed to help reduce maladaptive behaviors in ASD is through its repetitive characteristics and controlled flexibility through rhythm, melody and phrasing (Geretsegger et al., 2015). Another viewpoint posits that music alters behavior by modulating brainwave rates through entrainment (Orr et al., 1998). According to Orr et al. (1998), using metrically structured music alters alpha waves toward a relaxed and alert state ideal to reduce challenging behavior and increase focus. While studies reporting MI's efficiency to reduce problematic behaviors appear promising, Accordino et al. (2007) (2007) pointed out that such behavior is typically assessed within sessions, as opposed to accumulated effects as would be the case in a longitudinal approach. While it is useful to know whether MI can lead to reduced maladaptive behavior outside the therapeutic setting, it can still be considered an argument in favor of MI if clients engage in less disruptive and aggressive behavior during sessions.

As highlighted by Geretsegger et al. (2014), MIs are also used to enhance language skills in ASD. Rhythm and intonation in music can be used to help clients understand spoken language better (Accordino et al., 2007). Indeed, Lim (2010) points out that music perception is similar to speech perception in a variety of ways, e.g. pattern recognition and grouping of information into categories. Lim (2010) also notes that both music and speech are organized as frequency spectra that are perceptually identified as pitches. And as noted earlier, pitch perception is not impaired in ASD (LaGasse, 2014). Music might also reinforce language learning by virtue of its salience for the ASD population; Simpson, Keen, and Lamb (2015) note that a number of studies have found higher engagement in ASD children during a music-based intervention than the comparison condition. MI may also help non- and pre-verbal individuals with ASD to communicate without words (Geretsegger et al., 2015). Indeed, Buday (1995) argued that music can be used as a mediator to learn sign language in populations who suffer memory deficits or lack strategies to utilize their working memory efficiently.

A number of reviews and meta-analyses have studied MI in ASD. In a meta-analysis, Whipple (2004) examined outcomes of MI in ASD as compared to NMI. Although the individual studies used a variety of outcome measures, no differentiation between outcomes was made during quantitative analysis. The study concluded that MIs are generally effective in ASD children. Accordino et al. (2007) published a broad narrative review of MI in ASD that examined an array of outcomes, most notably social, communication and behavioral abnormalities. While noting the potential of MI in this population, the authors pointed out the limitations in past research, such as the predominant use of case studies as opposed to more generalizable study designs. In a systematic review with meta-analysis, Geretsegger et al. (2014) investigated improvements in social outcomes for MI compared to NMI and found significant effects, but found their analyses limited by the small sample sizes of the included studies, and the lack of consensus of outcome measures types between studies.

No systematic review to date has examined social, maladaptive and language measures with the aim of identifying specific effects of MI (compared to NMI) on these outcomes in children and adolescents with ASD. The present study attempted to go beyond past findings by using a rigorous method that focuses on bias reduction and considers different categories of behavioral outcome with respect to core dimensions of ASD. In an attempt to be as comprehensive as possible, a qualitative systematic review was performed for all available MI studies that met the inclusion criteria, followed by several quantitative meta-analyses by outcome type (social functioning, maladaptive behavior and language).

Objectives and Hypotheses

The present study set out to assess the efficacy of MI to address symptomatic deficiencies in ASD children and adolescents, by means of a systematic review and quantitative meta-analyses. All current literature was assessed in order to examine whether MIs present beneficial outcomes compared to NMIs in this population in terms of 1) enhancing social behavior outcomes, 2) reducing maladaptive behaviors, and 3) facilitating language skills.

In social behaviors, MI was expected to have a more positive effect, compared to NMI, on measures reflecting prosocial behaviors such as engagement, reciprocity and initiation of social interaction (Geretsegger et al., 2014; Kim et al., 2008). Similarly beneficial effects were also expected in social scales that include both prosocial and socially maladaptive behavior, such as the Social Responsiveness Scale.

In maladaptive behaviors, MI was expected to produce a more marked decrease in frequency and duration compared to NMI (Bhat & Srinivasan, 2013; Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004). These improvements were expected in terms of generally maladaptive behavior, such as sensory behaviors or negative affect, as well as socially maladaptive behavior, which occurs when participants show non-compliant behavior in social activities.

In language skills, MI was expected to result in improved performance in tasks such as receptive labeling as well as frequency of verbalizations compared to NMI, as hypothesized by a number of studies (e.g. Accordino et al., 2007; Geretsegger et al., 2014; Lim, 2010).

In addition to quantitatively testing these hypotheses by means of meta-analyses, the sample of studies also underwent systematic review in order to better describe current practices, assess sources of bias, and identify potential improvements for future work. This combined systematic review and meta-analyses aimed to provide an up-to-date evaluation of evidence of MI's benefits in ASD children and thus give recommendations for future clinical interventions and research about best practice in the domain.

Method

In order to answer the above questions, a qualitative systematic review, modelled after methods suggested by Cochrane Reviews (Higgins & Green, 2011) was conducted in order to describe music interventions and outcomes in included MI studies. In a first step, a comprehensive review of the available literature was conducted, in which studies were described in detail with respect to their design, sample size, intervention characteristics and control conditions. Studies' outcome measures were tabulated in accordance with the aims of this review (social outcomes, maladaptive behavior outcomes, and language outcomes), and study results for these outcome measures were compiled. Subsequently, quality assessments were performed on key characteristics of the studies and their respective interventions in order to identify sources of potential bias and lay bare the strengths and weaknesses of each selected study. Finally, after summarizing the findings of the included studies, effect sizes were calculated for all outcomes of interest in each study, and the efficacies of MI on social, maladaptive and language outcomes in children and adolescents with ASD were quantitatively tested using meta-analysis.

Implications

This procedure allows conclusions to be drawn about the current scientific state of MIs in ASD children. As highlighted by previous reviews and meta-analyses on MI's effect in ASD (Accordino et al., 2007; Geretsegger et al., 2014; Whipple, 2004), much of the research into MI in ASD has taken the approaches of case studies and qualitative descriptions, and generalizability of existing empirical research has been limited by small sample sizes and poor design in the past. The present thesis seeks to reevaluate the state of research in this area five years after Geretsegger et al. (2014) review. It serves as a pointer for practitioners as to the areas of functioning that MIs can best address in ASD. Additionally, this thesis informs future research by highlighting opportunities for more conclusive study methodologies that will further our understanding of MI's potential in ASD.

Abstract

Research increasingly suggests benefits of music interventions (MIs) in ASD. However, no systematic review employing meta-analysis has investigated MIs within key ASD symptoms of social functioning, maladaptive behaviors and language impairments. A systematic evaluation was performed across 17 studies comparing MIs with non-music interventions (NMIs) in ASD children. Meta-analyses of 11 studies supported effectiveness of MIs particularly for measures sensitive to social maladaptive behaviors. Comparisons further suggested benefits of MIs over NMIs for social outcomes, but not non-social maladaptive or language outcomes. Methodological issues were common in studies, such as small sample sizes, restricted durations of interventions, and underreporting of sample, matching, and attrition information. Recommendations are provided for future clinical interventions and research on MI in ASD.

Efficacy of Musical Interventions on Social, Maladaptive and Language Outcomes in Children with Autism Spectrum Disorder: A systematic review and meta-analysis

Introduction

General Introduction

A growing number of studies report that music interventions (MIs) have positive outcomes in clinical populations such as children and adolescents with autism spectrum disorder (ASD). Effectively, 12% of interventions in ASD employ music (Bhat & Srinivasan, 2013). With the increasing popularity of MIs, there is a growing need for rigorous assessment of their efficacy for symptom treatment. To our knowledge, no systematic review employing meta-analysis to date has investigated the efficacy of MIs across three of the predominant symptoms in ASD children and adolescents, specifically social functioning, maladaptive behaviors and language. Taking a comprehensive perspective on music interventions, the present study aims to inform both practitioners and researchers as to what favorable outcomes may be expected from MI in ASD children and adolescents. Accordingly, through systematic review and meta-analyses, all currently available research in ASD children and adolescents that compares MIs to non-musical interventions (NMI) is assessed in terms of core social, behavioral and communicative outcomes.

Music Interventions

MI uses music to achieve therapeutic goals, such as improving social function and reducing problematic behaviors (Cogo-Moreira et al., 2013). MI and music therapy (MT) are frequently used synonymously (e.g. Geretsegger et al., 2014; Jackson, 2003). However, MI can be implemented by any professional without any specific MT training (Maloy & Peterson, 2014), while MT is defined as the "clinical and evidence-based use of music interventions to accomplish individualized goals within a therapeutic relationship by a credentialed professional who has completed an approved music therapy program" (American Music Therapy

Association, 2013). Therefore, MT is also a MI, but not every MI qualifies as MT. The scope of the present study, in an effort to be as comprehensive as possible, will be inclusive of all types of MIs studied in ASD.

The use of music in MIs can take the form of a background or contingent stimulus, or as an interactive activity that engages patients to use their voice, movement and music instruments in a way that meets their therapeutic needs (Kim et al., 2009). Music is believed to have beneficial effects due to its multisensory characteristics, its communicational aspects and its enjoyability (Suthers & Niland, 2007). For example, in neurotypical children, some sources have reported that music has prosocial benefits (e.g. Kirschner & Tomasello, 2010; Miendlarzewska & Trost, 2014; Särkämö et al., 2013). MIs have been used effectively in various clinical populations such as dementia, schizophrenia and terminal illness (Tang & Vezeau, 2010). There is some evidence that MIs are also effective in ASD, a developmental disorder with rising prevalence (Accordino et al., 2007; Autism Speaks, 2018; Geretsegger et al., 2014). Especially taking into consideration that ASD children are said to enjoy music (Gebauer et al., 2014), it is not surprising that a rising number of interventions for this population employ music (Bhat & Srinivasan, 2013).

MI in ASD

Primary deficits in ASD include moderate to severe impairments in social-emotional functioning, maladaptive behaviors and communication and language skills. These deficits are present at an early age and cause significant impairment (American Psychiatric Association, 2013; LaGasse, 2017).

Many aspects of MI appear well suited to address both the strengths and core issues present in ASD. There is evidence that ASD children have enhanced auditory pitch perception and have stronger neural activations when exposed to songs compared to typically developing peers (LaGasse, 2014). Musical structure can bring a predictability that fulfills the need for anticipation in ASD, while allowing enough openness to address deficits such as joint attention, eye contact and turn-taking (Geretsegger et al., 2014; Geretsegger et al., 2015). Additionally, it has been noted that people with autism generally enjoy music (Gebauer et al., 2014) and even use it to manage their moods (Allen et al., 2009).

Building on these facilitating aspects of music, MI has the potential to promote improvements in eye contact, joint attention and turn taking, which are social behaviors ASD children frequently struggle with (Geretsegger et al., 2014). It has also been proposed that MIs can reduce maladaptive behaviors through their repetitive and predictable setup while allowing for flexibility through rhythm, melody and phrasing (Geretsegger et al., 2015). Another theory claims that music alters behavior by modulating brainwave rates via its metric structure (Orr et al., 1998). In terms of language impairments, fundamental similarities between music perception and speech perception provide avenues that could facilitate language learning such as pitch and intonation, rhythmicality, and auditory pattern processing (Accordino et al., 2007; Buday, 1995; Geretsegger et al., 2014; LaGasse, 2014; Lim, 2010).

Several previous reviews and meta-analyses have examined the potential of MI in ASD. These works report both a general support for MI's effectiveness and concerns about the limitations of existing research in this area. In a meta-analysis, Whipple (2004) examined outcomes of MI in ASD as compared to NMI. Although the individual studies used a variety of outcome measures, no differentiation between outcomes was made during quantitative analysis. The study concluded that MIs are generally effective in ASD children. Accordino et al. (2007) published a broad narrative review of MI in ASD that examined an array of outcomes, most notably social, communication and behavioral abnormalities. While noting the potential of MI in this population, the authors pointed out the limitations in past research, such as the predominant use of case studies as opposed to more generalizable study designs. In a systematic review with meta-analysis, Geretsegger et al. (2014) investigated improvements in social outcomes for MI compared to NMI and found significant effects, but found their analyses limited by the small sample sizes of the included studies, and the lack of consensus of outcome measures types between studies.

No systematic review or meta-analysis to date has examined social, maladaptive and language measures with the aim of identifying specific effects of MI (compared to NMI) on these outcomes in children and adolescents with ASD. The present study attempted to go beyond past findings by using a rigorous method that focuses on bias reduction and considers different categories of behavioral outcome with respect to core dimensions of ASD. In an attempt to be as comprehensive as possible, a qualitative systematic review was performed for all available MI studies that met the inclusion criteria, followed by several quantitative meta-analyses organized by outcome type (social functioning, maladaptive behavior and language).

Objectives and Hypotheses

The present study set out to assess the efficacy of MI to address symptomatic deficiencies in ASD children and adolescents, by means of a systematic review and quantitative meta-analyses. All current literature was assessed in order to examine whether MIs present beneficial outcomes compared to NMIs in this population in terms of 1) enhancing social behavior outcomes, 2) reducing maladaptive behaviors, and 3) facilitating language skills.

In social behaviors, MI was expected to have a more positive effect, compared to NMI, on measures reflecting prosocial behaviors such as engagement, reciprocity and initiation of social interaction (Geretsegger et al., 2014; Kim et al., 2008). Similarly beneficial effects were also expected in social scales that include both prosocial and socially maladaptive behavior, such as the Social Responsiveness Scale.

In maladaptive behaviors, MI was expected to produce a more marked decrease in frequency and duration compared to NMI (Bhat & Srinivasan, 2013; Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004). These improvements were expected in terms of generally maladaptive behavior, such as sensory behaviors or negative affect, as well as socially maladaptive behavior, which occurs when participants show non-compliant behavior in social activities.

In language skills, MI was expected to result in improved performance in tasks such as receptive labeling as well as frequency of verbalizations compared to NMI, as hypothesized by a number of studies (e.g. Accordino et al., 2007; Geretsegger et al., 2014; Lim, 2010).

In addition, the sample of studies underwent systematic review, modelled after methods suggested by Cochrane Reviews (Higgins & Green, 2011), in order to better describe current practices, assess sources of bias, and identify potential improvements for future work. This systematic review described studies in detail with respect to their design, sample size, intervention characteristics, control conditions, and quantitative outcomes. Subsequently, quality assessments were performed on key characteristics of the studies and their respective interventions in order to identify sources of potential bias and lay bare the strengths and weaknesses of each selected study.

Finally, after summarizing the findings of the included studies, effect sizes were calculated for all outcomes of interest in each study, and the efficacies of MI on social, maladaptive and language outcomes in children and adolescents with ASD were quantitatively tested using meta-analysis. Together, the combined systematic review and meta-analyses aimed to provide an up-to-date evaluation of evidence of MI's benefits in ASD children and thus give recommendations for future clinical interventions and research about best practice in the domain.

Methods

Rationale for study criteria

The present study investigated the efficacy of MI for social, maladaptive behavior and language outcome in ASD children and adolescents. In order to be included in this review, the study participants were expected to have a formal diagnosis of ASD or Asperger syndrome given by a clinician and be under 21 years of age.

This review was inclusive of any active MI, including MT. The presence of a formally trained music therapist was not necessary, in order to allow for inclusion of a wider array of MI in ASD. To be considered an intervention, the experimental condition was required to take place over the course of at least two days, as interventions in developmental disorders such as ASD are longitudinal in their nature. In order to evaluate the added benefit of music in interventions for ASD, this review selected only studies that compared an active music intervention condition with a non-music intervention using measures of social functioning, autistic behaviors such as repetitive or maladaptive behavior, or language outcomes. MIs used music actively by letting the children interact with music instruments, sing and/or dance, or presented an active listening task to the participants. MIs with a more passive nature, such as auditory integration therapy (AIT), were therefore excluded. For inclusion in the meta-analyses, the control condition had to be an intervention that used a non-music approach; wait-listed controls or different forms of MI were not included as control conditions.

Studies were required to follow a randomized controlled trial (RCT) design. This means that participant allocation to the conditions (experimental or control) must be determined randomly or pseudo-randomly, in order to promote equal distribution of population characteristics such as age, gender, socio-economic status (SES) or symptom severity among groups (Higgins & Green, 2011). This helps avoid sources of bias that could confound the study findings as well as the present systematic review and meta-analyses.

Study outcomes were either measured using standardized tests and questionnaires such as the Social Responsiveness Scale (SRS; Constantino & Gruber, 2012) or Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997), or behavioral measures and video-coding of behaviors specific for either of these aims, such as frequency of initiation of engagement or frequency of interested affect. For meta-analyses, the study measures were categorized to facilitate testing the specificity of MI effects on social functioning, maladaptive behavior and language outcomes. The measure categorization was thus based on these three a priori aims as well as distinctions that were observed in the course of reviewing and coding the studies. Social functioning measures either assessed social behavior in a positive way, measuring strictly presence of prosocial behavior (e.g. , Joint Attention Test, JTAT; Bean & Eigsti, 2012) or combined presence and lack of prosocial behavior (e.g. the SRS), or assessed only social maladaptive behavior (e.g., Vineland Adaptive Behaviour Scale, VABS; Sparrow & Cicchetti, 1985). Maladaptive behavior overall fell into the latter category of social maladaptive behavior, or assessed non-social aspects of maladaptive behavior (e.g., frequency of sensory behavior). In language outcomes, there was a distinction between measures that assessed correct reproduction of words (e.g., accuracy of receptive labeling) and those that assessed language in a more practical social setting (e.g., frequency of verbalizations); however, there were not enough studies using language outcomes to justify this division in quantitative analysis. Following from these distinctions, the set of measure categories was “prosocial”, “mixed social”, “social maladaptive”, “non-social maladaptive”, and “language”. Analyses were performed for these individual categories and several groupings of categories as described further in the quantitative meta-analyzes methods.

A separate analysis distinction was made based on a study's design and the question that it effectively posed. Some outcomes were measured in terms of longitudinal changes between the beginning and end of the intervention period, whereas others were assessed within sessions. Studies using the longitudinal approach all followed the pre-post-control (PPC) approach, in which two independent groups differing in intervention condition (i.e., here, MI and NMI) are tested on outcomes before and after the intervention period (Morris, 2008). Studies following the within-session approach used either independent groups or a crossover design. As these two paradigms of measuring intervention outcomes answer two related but distinct types of questions — specifically, whether MIs improve behaviors during intervention sessions in within-session assessments, and whether MIs lead to different outcomes at the end of interventions compared to the beginning in longitudinal evaluations — these designs were considered separately in the present meta-analyses.

Literature search

A thorough literature search was conducted in the online databases Web Of Science, Scopus, PsychInfo, ERIC, CINAHL and PubMed to identify relevant studies for the systematic review and meta-analyses. Papers published between 1995 and 2018 were queried in each of these databases using the keywords ("ASD", "autism", "asperger", "autism spectrum disorder") AND ("music therapy", "music intervention" and "music"). The keywords were kept general to reduce the risk of omitting studies relevant to the present aims. Additionally, hand searches were conducted by looking up selected papers' references in order to find additional papers that may not have been identified in the database queries.

Inclusion process

All studies found in the database queries and hand searches were imported into EndNote X8.2, where in a first step, duplicates were removed. Following this, title and abstract screening was conducted to remove all studies that did not include a sample of children or adolescents with ASD. In a next step, through initial full-text examination, studies were removed if they did not investigate MI. Finally, studies were removed if they did not use quantitative methods, failed to follow a RCT design, were published in a language other than English, French or German, or were not published in peer-reviewed journals.

The remaining studies were entered into a coding spreadsheet. At this coding stage, a more rigorous set of inclusion criteria was applied in order to select studies for the systematic review and meta-analyses (note that the meta-analyses required additional criteria mentioned in the second list):

- A minimum of 10 participants
- At least one outcome measure in social functioning, maladaptive behavior, or language performance
- Absence of adult participants (age <21 years)
- Presence of a control condition
- A design conforming to the definitions of either "longitudinal" or "within-session" as described earlier
- A minimum intervention duration of two days
- An active MI condition (i.e., that either employed active interaction with music instruments or active listening to musical stimuli by the participants)
- For inclusion in the quantitative analyses, there were several additional requirements:
- Numeric results sufficient to calculate effect sizes, such as a pre-calculated standardized mean difference (e.g., Cohen's d), means with standard deviations, or means with confidence intervals.
- Presence of a NMI control condition (this excluded wait-listed controls or a comparison between different MIs).

The coding sheet tabulated data about design type (either longitudinal or within-session), role of music (active listening or active production), measure category ("prosocial", "mixed social", "social maladaptive", "non-social maladaptive" or "language"), sample size, age mean and range of participants, duration and number of sessions, duration of intervention and total intervention hours, as well as the mean, standard deviation and confidence intervals for study outcomes. Coding was performed separately by two researchers (NR and KJ), and interrater reliability was 94.6%.

Quality assessment

In a further effort to assess the impact of bias in the reviewed studies, a rigorous quality assessment based on the CONSORT criteria (Moher, Liberati, Tetzlaff, & Altman, 2009) was conducted by two researchers. Specific information in this assessment included studies' use of ASD diagnostic instruments, whether sample size exceeded 20 participants per group, adherence to randomization and blinding, reported participant information (age, gender, IQ, symptom severity, SES of parents), matching criteria, matching status, attrition bias, study rationale, quality of intervention, outcome measure quality, and reporting bias. Each criterion was rated with a grade of 0 (high risk of bias), 1 (unknown risk of bias), or 2 (low risk of bias). This quality assessment provided a consistent approach to grade the quality of the included studies by flagging possible sources of bias. Despite the fact that some studies failed to specify their methods for randomization and blinding, a decision was made to retain these studies in the quantitative meta-analyses due to the limited number of studies available.

Addressing risk of bias

Risk of bias was assessed according to information provided about diagnostic criteria, allocation randomization, blinding of those who evaluated intervention outcomes, participants and their caregivers, reporting of baseline characteristics per group, matching of participants according to age, gender, IQ, symptom severity and SES, handling of missing data and quality of data reporting.

Assessing quality of interventions

The quality assessment further looked into the quality of music interventions presented in each study. Intervention quality was rated according to intervention duration, stated reasoning for the type of intervention given, quality of intervention protocol, adequacy of control intervention, adequacy of measures used to assess intervention outcomes, qualification of the person giving the intervention, and consistent adherence to the intervention throughout the study.

Quantitative meta-analysis

In the quantitative meta-analyses, effect sizes (standardized mean differences, SMDs) were calculated for each selected measure in each study, and then each analysis of interest was conducted in accordance with standard meta-analytic guidelines (Lipsey & Wilson, 2001). The meta-analyses were organized based on groupings of measures according to this review's aims, as well as by study design (longitudinal or within-session, as described above). The measure groupings comprised each of the measure categories ("prosocial", "mixed social", "social maladaptive", "non-social maladaptive", and "language") as well as an overall "social" grouping (combining prosocial, mixed social, and social maladaptive), an overall "maladaptive" grouping (combining social maladaptive and non-social maladaptive, and an "overall" grouping that included all measure categories.

Several different equations were used to calculate the SMDs and their variances, in order to properly account for covariance arising from multiple measurements per participant, and to reduce small-sample biases. Longitudinal studies all reported pre- and post-intervention measures for the MI and NMI groups. SMDs representing the difference in pre-post change between intervention groups were calculated using dppc2 (Morris, 2008; equation 8). The variance of this SMD was calculated according to Morris' equation 25 (2008). The SMD variance equation makes use of a correlation value ρ that represents the degree of correlation between multiple measurements within a participant (i.e., pre-intervention and post-intervention). When the necessary values were reported by a study, these correlations were calculated based on equation 4.27 of Borenstein, Hedges, Higgins, and Rothstein (2011). Correlations were calculable for total of 11 longitudinal effect sizes, whereas for 7 effect sizes the necessary values were not available. In the case of unavailable correlation values, the median of the known correlation values ($r = .753$) was imputed. In order to gauge the sensitivity of each analysis to this imputation, supplemental analyses were run using the minimum ($r = .272$) and maximum ($r = .999$) extremes of the known correlation values. In cases where imputing at the minimum or maximum value changed a meta-analysis result (from significant at $p < .05$ to non-significant, or vice-versa), this is noted in the results text. Results were generally quite robust across these imputed r values (c.f. Table 3).

Within-session studies either compared measurements from independent MI and NMI groups, or used a crossover approach in which the MI and NMI conditions were measured in all participants. For studies having independent groups, SMDs were calculated incorporating Hedges' correction for small sample bias (Hedges 1981) using equation 4.23 in Borenstein et al. (2011). SMD variance was calculated using equation 4.24 from Borenstein et al. (2011). For studies using a crossover design, SMDs and their variances were calculated using the equations provided in section 16.4.6.2 of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011). For crossover studies, the SMD and variance equations make use of a correlation value ρ that represents the degree of correlation between multiple measurements within a participant (i.e., MI and NMI). Only one crossover study (Lim & Draper, 2011) reported the necessary values to calculate a correlation value, and using the calculated value resulted in an extremely high weighting of this study in any meta-analysis in which it was included, such that the meta-analysis result essentially represented only this study. As this was undesirable, the pragmatic decision was made to impute the correlation value at $r = .500$ for this study as with the other crossover studies. In order to gauge the sensitivity of each analysis to this imputation, supplemental analyses were run using extreme minimal ($r = .000$) and maximal ($r = .999$) imputed correlation values. In cases where imputing at the minimal or maximal value changed a meta-analysis result (from significant at $p < .05$ to non-significant, or vice-versa), this is noted in the results text. Results were generally quite robust across these imputed r values (c.f. Table 3)

To ensure consistency of effect size direction across different measures, effect sizes representing a beneficial outcome (e.g., decreased maladaptive behavior, or increased prosocial behavior) were given a positive sign, and effects representing an adverse outcome were given a negative sign.

Each meta-analysis (separately by measure grouping, study design, and correlation imputation level, in all cases where there were at least 2 studies to analyze) was calculated in R software version 3.5.3 (R Core Team, 2019) using the "rma.mv" function in the metafor package version 2.0 (Viechtbauer, 2010). The rma.mv function performs a random-effects, multi-level analysis that can account for correlated sampling errors when multiple effects are included per study. The random-effects approach accounts for variation across studies and is recommended when studies vary in their samples or methodology (Borenstein et al., 2011). Moreover, random effects provide for greater control for differences in sample size when estimating effect sizes (Borenstein et al., 2011). Between-study heterogeneity was estimated using the restricted maximum likelihood (REML) method and confidence intervals of effect sizes were estimated using the Q-profile method, following the recommendations of Veroniki et al. (2016). Confidence intervals are reported at the 95% level, p values were considered significant at values $< .05$, and in consideration of the low power of generally small sample sizes in the reviewed studies, non-significant trends at $p < .10$ are also noted.

Results

Outcome of literature search

An extensive electronic literature search was conducted in September 2018 and yielded 993 results after duplicates were removed. Of 740 studies deemed relevant from their titles and abstract, 210 were excluded after full-text searches because they did not fit into an RCT design or were of qualitative methodology. Six studies were excluded because they were in a language other than English, French or German and there were no means to translate them. During coding, another 43 studies were excluded because they had an insufficient number of participants, lacked a non-music control condition, were conducted within just one day or did not match the aims of the present review upon closer inspection. A total of 17 studies remained for qualitative assessment after these exclusions. For the quantitative meta-analyses, six studies could not be included because they did not report data necessary for statistical analysis (Katagiri, 2009; Sandiford, Mainess, & Daher, 2012), used a waitlisted control design (Ghasemtabar et al., 2015) rather than including a non-music control intervention, or used the same participants as another included study deemed more relevant for our purposes (Kim et al., 2008; Srinivasan et al., 2016a; Srinivasan et al., 2016b), leaving 11 studies for meta-analysis. Figure 1 shows the selection process in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Moher et al., 2009).

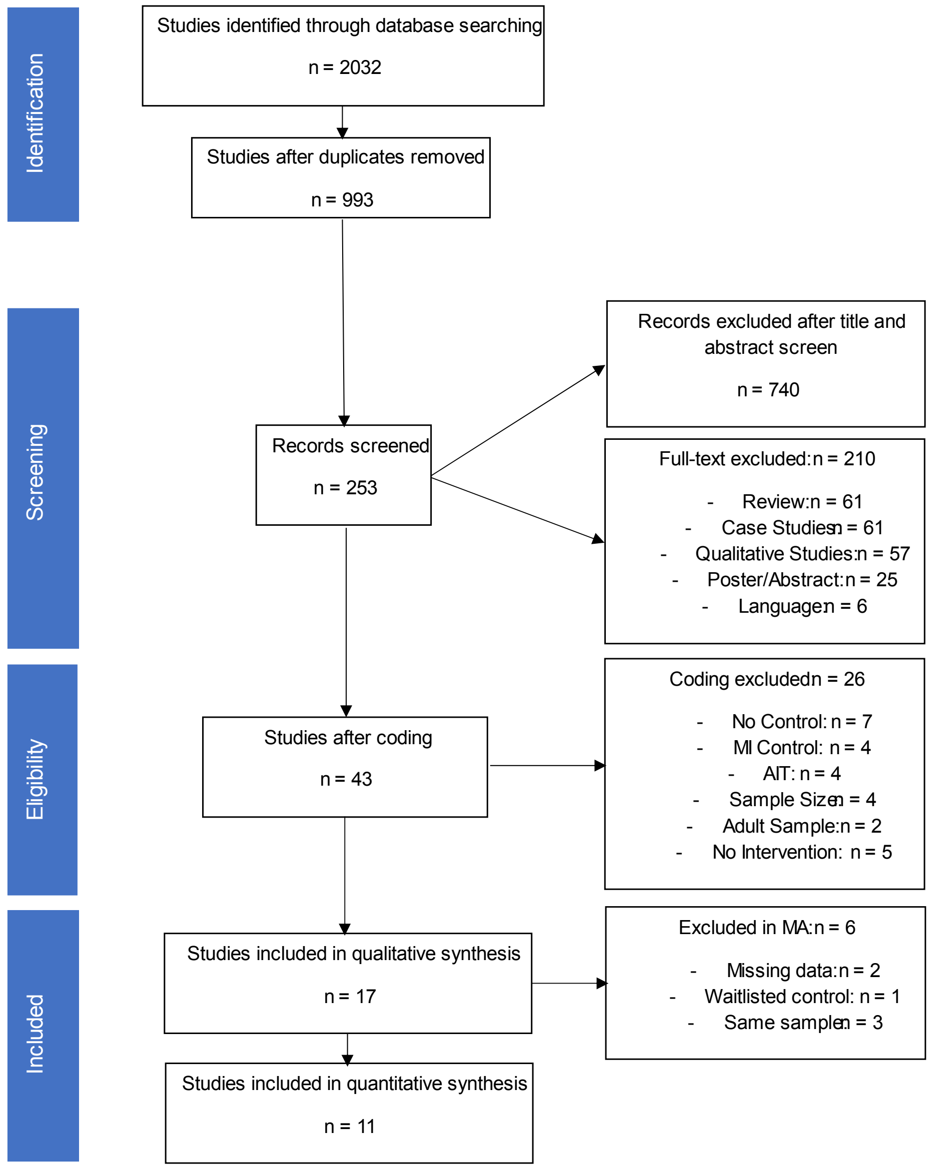


Figure 1: PRISMA flow chart of the literature search

Included studies

Table 1 summarizes the included studies for this review and meta-analysis, and lists the outcome measures selected for this analysis. Although the database searches were inclusive of studies from 1995 to 2018, the publication years of the studies that met criteria for the systematic review ranged from 2007 to 2018. It is to be noted that with the exception of four studies (Ghasemtabar et al., 2015; Katagiri, 2009; Kim et al., 2008, 2009), there was a notable greater representation of Western countries, with nine studies conducted in the United States, two in the United Kingdom, one in Australia and one in Canada. Interestingly, one study was conducted in nine countries (Bieleninik et al., 2017).

Study designs

All studies assessed the effect of MI over a minimum of two days. In the present analyses, a distinction was made between "longitudinal" and "within-session", and this classification was related both to the design of a study and how the behavioral measurements were made. Longitudinal studies assessed behavior before and after completion of an intervention, in order to compare the pre-post change of a MI with a NMI. Twelve studies employed such a longitudinal design (Bieleninik et al., 2017; Crawford et al., 2017; Gattino et al., 2011; Ghasemtabar et al., 2015; Katagiri, 2009; Lim, 2010; Sandiford et al., 2012; Schwartzberg & Silverman, 2013; Sharda et al., 2018; Srinivasan et al., 2016a; Srinivasan, Park, Neelly, & Bhat, 2015; Thompson, McFerran, & Gold, 2013). Three of these longitudinal studies also had follow-up investigations at two months (Ghasemtabar et al., 2015) or 12 months (Bieleninik et al., 2017; Crawford et al., 2017). Within-session studies investigated behaviors during the sessions, rather than before and after the intervention protocol as a whole, and therefore did not lend themselves to assessing pre-post changes. Seven studies were of within-session design, and of these, three used an independent-groups approach (Katagiri, 2009; Srinivasan et al., 2015; Srinivasan et al., 2016a), and four used a crossover approach in which all participants completed both the music and control intervention (Kim et al., 2008, 2009; Lim & Draper, 2011; Simpson et al., 2015).

Table 1: Summary of studies included in the systematic review and meta-analyses. Asterisks mark studies included in the quantitative meta-analyses. Measure categories for the meta-analyses are indicated in parentheses in the Outcomes column. PPC: Pre-post-control; POWC: Post-only with control.

* marks studies included in meta-analyses

[1] Bieleninik et al. (2017) was conducted in Australia, Austria, Brazil, Israel, Italy, Korea, Norway, United Kingdom, United States.

Study ID	**Year**	**Countries**	**Design**	***n***	**Intervention Hours**	**Outcomes**
*Bieleninik et al.[1]	2017	Various[1]	Longitudinal (PPC)	314	33	**Social (mixed social):** ADOS-social effects: p = .88, no significant improvement SRS total: no significant improvement
*Crawford et al.	2017	United Kingdom	Longitudinal (PPC)	64	33	**Social (mixed social):** ADOS-social effects: no significant improvement SRS total: no significant improvement
*Gattino et al.	2011	United Kingdom	Longitudinal (PPC)	24	8	**Social (pro-social):** CARS-BR: no significant difference (verbal communication p=.50; non-verbal communication p=.35; social communication p=.34)
Ghasemtabar et al.	2015	Iran	Longitudinal (PPC)	27	12	**Social:** SSRS: significant improvement in MT group

Katagiri	2009	Japan	Longitudinal (POWC)	12	4	**Social:** Understanding of emotions: most significant gains in music condition, *p*=.01
Kim et al.	2008	Korea	Within Session (Crossover)	10	6	**Social:** ESCS: Music was significantly more effective than play therapy, *p*=.01 **Maladaptive Behaviors:** PDDBI: no significant effect for condition
*Kim et al.	2009	Korea	Within Session (Crossover)	10	6	**Social (pro-social):** significant effect in favor of music found for joy, emotional synchronicity; initiation of engagement and compliant response, all *p*<.001 **Maladaptive Behaviors (social maladaptive):** less instances of no response in music condition, *p<.001*
*Lim	2010	United States	Longitudinal (PPC)	36	1	**Language (language):** VPES: both music (*d*=1.275) and speech condition (d = 1.141) had a large effect size
*Lim & Draper	2011	United States	Within Session (Crossover)	22	.7	**Language (language):** VPES: significant effect for both, but no significant difference between music and speech condition

Sandiford et al.	2013	United States	Longitudinal (POWC)	12	3.75	**Language:** Significant increase in verbal attempts (*p*<.001), correct words (*p*=.04) and imitative attempt (*p*=.01) in both groups, but difference in treatment effect was not significant (*p*=.08)
*Schwartzberg & Silverman	2013	United States	Longitudinal (PPC)	30	7	**Social (pro-social):** ASSP (SP and SR): there was no significant difference between conditions **Maladaptive Behavior (social maladaptive):** ASSP (DSB): there was no significant difference between conditions
*Sharda et al.	2018	Canada	Longitudinal (PPC)	51	7.5	**Social (pro-social):** CCC-2: significant improvements in communication outcomes; **Social (mixed social):** SRS-II: no group differences were detected **Maladaptive Behaviors (social maladaptive):** VABS: significant decreases in maladaptive behavior in both groups were found **Language (language):** PPVT-4: no group differences were detected

*Simpson et al.	2013	Australia	Within Session (Crossover)	22	1.25	**Social (pro-social):** level of engagement: a significant main effect for condition was found, *p*=.04 **Maladaptive Behaviors (social maladaptive):** challenging behavior: no significant difference between sung and spoken condition, *p*=.065 **Language (language):** correct responses were strongly correlated with level of engagement (*p*<.01) but not with challenging behavior (*p*=.06 – *p*=16.0)
*Srinivasan et al.	2015	United States	Within Session (Independent-groups)	36	24	**Social (pro-social):** Positive affect: the rhythm group had significantly greater positive affect than the comparison group in the mid-sessions, *p*=.006. Interested affect: the duration of interested affect was greater in the comparison group in all sessions, *p*<.05. But the rhythm group experienced a significant increase in interested affect, *p*<.002. **Maladaptive Behaviors (social maladaptive):** Negative affect: the rhythm group showed greater negative affect in all sessions, *p*<.05, but showed a

						more marked decrease than the comparison group, *p*<.002. Negative behaviors: the rhythm group showed more negative behaviors than the comparison in the early sessions only, *p*<.05. **Maladaptive Behaviors (non-social maladaptive):** Sensory behaviors: the comparison group showed more sensory behaviors than the rhythm group (*p*<.05). Stereotyped behaviors: there were no significant differences between groups
Srinivasan et al.	2016a	United States	Longitudinal (PPC)	36	24	**Social:** JTAT: the rhythm group showed significant improvements over time (*p*=.005), but so did the comparison group (*p*=.004). There were no between-group differences. **Maladaptive Behaviors:** Social attention: the rhythm group showed more attention to social partners and elsewhere than the comparison group (*p*<.01). The comparison group showed more attention to other objects than the

						rhythm group (p<.001). There were no significant changes in social attention over time.
Srinivasan et al.	2016b	United States	Longitudinal (PPC)	36	24	**Social:** Social verbalization: The rhythm group showed as significant increase in verbalizations toward the trainer over time (p=.02), whereas the comparison group showed no changes over time. There are no significant differences in social verbalizations between the rhythm and comparison group. **Maladaptive Behaviors:** Social bids: the rhythm group showed significant increases in total word count from early to late (p<.03) and mid to late sessions (p<.03)
*Thompson et al.	2013	Australia	Longitudinal	23	9.9	**Social (pro-social):** VSEEC: there was a statistically significant effect for MI, p<.001, with a very large effect size, d=1.96. **Social (mixed social):** SRS-PS: there was no significant difference for SRS-PS. **Language (language):**

						MBCDI-W&G: both intervention groups made improvements in parent-reported speech and language skills, but with no significant effect for treatment.

Study sizes

This study's sample consists of a total of 683 participants. The largest study took place in nine different countries (Bieleninik et al., 2017) and analyzed a total of 314 participants. Crawford et al. (2017) employed the same design as Bieleninik et al. (2017), but this review selected their data reported from the NHIL-funded arm of the study, consisting of 64 participants recruited from the United Kingdom. Sharda et al. (2018) had 51 participants. Ten studies ranged between 50 and 20 participants (Gattino et al., 2011; Ghasemtabar et al., 2015; Lim, 2010; Lim & Draper, 2011; Schwartzberg & Silverman, 2013; Simpson et al., 2015; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b; Thompson et al., 2013). Four studies had relatively small samples between 10 and 20 participants (Katagiri, 2009; Kim et al., 2008, 2009; Sandiford et al., 2012). The three studies by Srinivasan (2015; 2016a; 2016b) and the two studies by Kim et al. (2008, 2009) used the same sample.

Intervention durations and intensities

The studies varied greatly in intervention duration and intensity. Table 2 shows a summary of all intervention durations, number of sessions and session durations. The longest interventions were reported by Bieleninik et al. (2017) and Crawford et al. (2017) and lasted five months. Gattino et al. (2011) and Thompson et al. (2013)'s interventions lasted 16 weeks. Six studies lasted between 12 and 6 weeks (Ghasemtabar et al., 2015; Kim et al., 2008, 2009; Sandiford et al., 2012; Sharda et al., 2018; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b). Three studies lasted four to two weeks (Katagiri, 2009; Lim, 2010; Lim & Draper, 2011; Simpson et al., 2015). The remaining two studies lasted a week or less (Lim, 2010; Schwartzberg & Silverman, 2013).

Most sessions lasted between 30 and 45 minutes. The sessions offered by Ghasemtabar et al. (2015) lasted an hour, whereas the interventions by and Simpson et al. (2015) lasted less than 10 minutes.

In terms of intensity, four studies offered weekly interventions (Gattino et al., 2011; Kim et al., 2008, 2009; Thompson et al., 2013), two had two sessions a week (Ghasemtabar et al., 2015; Katagiri, 2009), Lim and Draper (2011), Sandiford et al. (2012) and Srinivasan et al. (2015; 2016a; 2016b) offered three weekly sessions. Lim (2010)'s intervention only lasted three days in total, but started with two training sessions on the first day, followed by six short sessions on the next. Simpson et al. (2015)'s study featured 15 sessions, but it is not clear over what time period. Schwartzberg and Silverman (2013) offered daily interventions during a weeklong camp for autistic youth. Bieleninik et al. (2017) and Crawford et al. (2017) had an intensive group receiving three weekly sessions, whereas another group received one weekly session.

Participants

Age of participants

Participants' ages ranged from three to 21 years, as set by this review's inclusion criteria, though most studies worked with younger children aged between three and seven (Anderson, Kish, & Cornell, 1980; Bieleninik et al., 2017; Crawford et al., 2017; Kim et al., 2008, 2009; Lim, 2010; Lim & Draper, 2011; Sandiford et al., 2012; Thompson et al., 2013). Others worked with older children aged six to 12 (Gattino et al., 2011; Ghasemtabar et al., 2015; Sharda et al., 2018). Some studies included larger age ranges from three to nine (Simpson (Simpson et al., 2015), five to 12 (Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b), or nine to 21 (Schwartzberg & Silverman, 2013).

Table 2: Intervention durations and intensities

* marks studies included in meta-analyses

Study ID	**Intervention Duration**	**Number Sessions**	**Total Intervention Hours**	**Session Intensity**
*Bieleninik et al. 2017	26 weeks	105	33	30 mins
*Crawford et al. 2017	26 weeks	105	33	30 mins
*Gattino et al. 2011	16 weeks	16	8	30 mins
Ghasemtabar et al. 2015	6 weeks	12	12	60 mins
Katagiri 2009	4 weeks	8	4	30 mins
Kim et al. 2008	12 weeks	12	6	30 mins
*Kim et al. 2009	12 weeks	12	6	30 mins
*Lim 2010	1 week (3 days)	6	1	20 mins
*Lim & Draper 2011	2 weeks	6	.7	> 10 mins
Sandiford et al. 2013	5 weeks	5	3.75	45 mins
*Schwartzberg & Silverman 2013	1 week	7	7	50 mins
*Sharda et al. 2018	8 – 12 weeks	8 – 12	7.5	45 mins
*Simpson et al. 2013	? weeks (15 days)	15	1.25	> 10 mins
*Srinivasan et al. 2015	8 weeks	32	24	45 mins
Srinivasan et al. 2016a	8 weeks	32	24	45 mins
Srinivasan et al. 2016b	8 weeks	32	24	45 mins
*Thompson et al. 2013	16 weeks	16	9.9	45 mins

Gender of participants

The majority of participants in the selected studies were male, with three studies using males exclusively (Gattino et al., 2011; Kim et al., 2008, 2009). Four studies neglected including information on participants' gender (Katagiri, 2009; Lim, 2010; Simpson et al., 2015; Thompson et al., 2013). Ghasemtabar et al. (2015) was the only study with an equivalent balance between females and males, with 13 girls compared to 14 boys. For four studies, the percentage of females was below 25% (Bieleninik et al., 2017; Crawford et al., 2017; Lim & Draper, 2011; Sharda et al., 2018). For another four, it was below 10% (Schwartzberg & Silverman, 2013; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b). These results are hardly surprising, considering the male to female diagnostic ratio is 3:1 (Loomes, Hull, & Mandy, 2017).

Diagnosis confirmation

Most studies confirmed the diagnosis of ASD using standardized measures such as the Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 2003), Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994) or the Childhood Autism Rating Scale (CARS; Schopler, Reichler, DeVellis, & Daly, 1980) (Bieleninik et al., 2017; Crawford et al., 2017; Gattino et al., 2011; Ghasemtabar et al., 2015; Kim et al., 2008, 2009; Lim, 2010; Sandiford et al., 2012; Sharda et al., 2018; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b). Other studies reported that participants held a former diagnosis of ASD (Lim & Draper, 2011; Thompson et al., 2013). Simpson et al. (2015) additionally consulted the Social Communication Questionnaire (SCQ; Berument, Rutter, Lord, Pickles, & Bailey, 1999), which as a social measure, is not specifically designed for diagnosis of ASD per se. Schwartzberg and Silverman (2013) only mentioned that their participants were recruited from a summer camp for ASD children. Katagiri (2009) did not specify how the diagnosis was confirmed.

Intervention settings

Interventions most frequently took place in outpatient therapy centers (Gattino et al., 2011; Ghasemtabar et al., 2015; Kim et al., 2008, 2009; Lim, 2010; Sandiford et al., 2012) or in schools (Simpson et al., 2015) or at home (Thompson et al., 2013), or a combination of these settings (Bieleninik et al., 2017; Crawford et al., 2017; Katagiri, 2009). Schwartzberg and Silverman (2013)'s intervention took place during a summer camp for ASD children. Five papers (Lim & Draper, 2011; Sharda et al., 2018; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b) did not report the setting of their intervention.

Interventions

Music Interventions

The majority of studies used a one-on-one setting for music interventions, with the exception of Kim et al. (2008, 2009) and Schwartzberg and Silverman (2013)'s interventions. Thompson et al. (2013) used a family-centered approach, actively encouraging active parental participation. Srinivasan et al. (2015; 2016a; 2016b) included an adult model in addition to the dyadic child/interventionist setup. Nearly all interventions were administered by experienced therapists or therapy students. Some studies did not report who administered the MIs (Ghasemtabar et al., 2015; Katagiri, 2009; Simpson et al., 2015).

Five studies used an improvisational or semi-structured music therapy approach (Bieleninik et al., 2017; Crawford et al., 2017; Kim et al., 2008, 2009; Sharda et al., 2018). This type of therapy allows the therapist and patient to join in spontaneous music co-creation involving singing, use of music instruments and dancing. It demands the therapist to attune to the child's momentary needs and interests to facilitate joint attention.

Four studies used a more structured approach (Schwartzberg & Silverman, 2013; Srinivasan et al., 2015; Srinivasan et al., 2016a; Srinivasan et al., 2016b) in which there was a clear progression from a welcoming sequence over to different types of musical activities involving singing, using instruments and different movements, and finishing with a goodbye-song.

It is important to note that Schwartzberg and Silverman (2013)'s study used music therapy as a frame for their actual intervention, which employs social stories that are either sung or spoken. Simpson et al. (2015) also compared sung and spoken conditions. The sung condition used a child-friendly melody to facilitate vocabulary learning through a play-song activity.

Gattino et al. (2011) used relational music therapy, which resembles a hybrid form between structural and improvisational music therapy, using both improvising and music games, composing and singing to reach therapeutic goals, while attuning to the patient's needs. Ghasemtabar et al. (2015)'s intervention was similar, but used Orff-schulwerk instruments (xylophone, triangle, bells, metallophone, xylophone, maracas, castanets, tambourine and woodblock).

Lim (2010) used developmental speech and language learning through music (DSLM), which uses songs to enhance vocabulary learning. It specifically employs typical musical elements such as pitch and rhythm in accordance with the Gestalt perceptual law of good continuation to prompt the patient to finish the sequence with the target word. Lim and Draper (2011) expanded on that concept by incorporating music in an applied behavior analysis verbal behavior (ABA VB) intervention. The verbal instruction was sung instead of spoken using child-appropriate songs.

Sandiford et al. (2012) tested melodic based communication training (MBCT), which uses a specific melody per target work in conjunction with pre-recorded instrumentals to accompany the melody and rhythmic clapping to learn the names of stimulus items.

Katagiri (2009) used two different music interventions to help children gain understanding about emotions. In the background music condition, melodies were played that were supposed to match the target emotion while the instruction was read to the participants. In the sung song condition, a song was sung along the melody, with lyrics matching the target emotion.

Control interventions

Three studies had control groups who received enhanced standard care (Bieleninik et al., 2017; Crawford et al., 2017; Thompson et al., 2013). Gattino et al. (2011) and Srinivasan et al. (2015; 2016a; 2016b) offered activities that are not specifically enhanced standard care, but routinely used in ASD children, such as seatwork activities. It is to be noted that Srinivasan et al. (2015; 2016a; 2016b) also used a second condition, in which a robot replaced the therapist and no music was used for movement activities. Lim and Draper (2011) used Applied Behavior Analysis Verbal Behavior (ABA VB) in their control group, omitting the music component. Sandiford et al. (2012) also left out the music component in their control condition. Lim (2010), Schwartzberg and Silverman (2013) and Simpson et al. (2015) compared a sung with a spoken condition, so the control condition resembled the experimental condition but used spoken instructions without musical accompaniment instead. Kim et al. (2008, 2009) used therapeutic play sessions as a comparison intervention. Katagiri (2009) used both a no contact condition, in which controls did not receive any intervention, and a non-music control, in which only verbal instructions were used to teach emotion understanding. Finally, Ghasemtabar et al. (2015) only used waitlisted controls in their study.

Outcome measures

The literature review identified a total of 14 studies that measured social outcomes, eight that assessed maladaptive behaviors, and six that evaluated language in concordance with the present aims.

Outcomes in social measures

Fourteen of the analyzed studies examined social effects of music interventions. Bieleninik et al. (2017) and Crawford et al. (2017) used the social effects subscale of the ADOS, a diagnostic tool for autism that is evaluated by clinicians. Gattino et al. (2011) used the CARS adapted for use in Brazil (CARS-BR; Pereira, 2007), which is another diagnostic tool rated by clinicians. Three studies (Bieleninik et al., 2017; Crawford et al., 2017; Sharda et al., 2018) used the SRS, which is administered to participants' parents. Sharda et al. (2018) additionally used the Children's Communication Checklist (CCC-2; Bishop, 2013), which is also given to parents or caregivers. Ghasemtabar et al. (2015) evaluated social behavior using the parental version of the Social Skills Rating Scale (SSRS; Gresham & Elliott, 1990). Kim et al. (2008) used the Early Social Communication Scale (ESCS; Seibert, Hogan, & Mundy, 1982), which assesses initiation of joint attention and responses to joint attention bids, typically in play therapy settings. In their 2013 study, Schwartzberg and Silverman used the Autism Social Skills Profile (ASSP; Bellini & Hopf, 2007) to assess generalization of targeted social skills, which are divided into Social Reciprocity (SR), Social Participation (SP) and Detrimental Social Behaviors (DSB). The former two measures were included in the present review's social outcomes aim, whereas the last one was considered to be related with maladaptive behaviors. Finally, Thompson et al. (2013) used the Vineland Social-Emotional Early Childhood Scales (VSEEC; Sparrow, Balla, & Cicchetti, 1998), which is administered via a semi-structured interview with participants' parents. Srinivasan et al. (2016a) used a modified version of the JTAT, comprised of naturalistic verbal and non-verbal social bids.

Other studies used non-standardized coding of observed or videotaped behaviors. Katagiri (2009) coded for correct identification of displayed emotions on pictures and photographs. In their 2009 paper, Kim et al. coded for duration and frequency of joy, emotional synchronicity, initiation of engagement and compliant response to therapist's bids. Simpson et al. (2015) also coded for level of engagement observed in participants during intervention sessions. Srinivasan et al. (2015) coded for positive, negative and interested affect displayed in sessions, and in their second 2016 paper, Srinivasan et al. frequency and duration of social verbalizations by participants.

Summary of findings in social measures

As reported in Table 1, the majority of studies report a significant effect in favor of MI. Significant effects for MI were found in the SSRS (Ghasemtabar et al., 2015), ESCS (Kim, Wigram & Gold, 2008) and VSEEC (Thompson et al., 2013). Sharda et al. (2018) found significant improvements on the CCC-2, but not on the SRS-II. The JTAT showed significant improvements in both the MI and comparison group, but there was no significant difference between the two. Studies did not find significant changes on ADOS social effect scores (Bieleninik et al., 2017; Crawford et al., 2017) or the SRS (Bieleninik et al., 2017; Crawford et al., 2017; Sharda et al., 2018), nor on the CARS-BR (Gattino et al., 2011) and the ASSP (Schwartzberg & Silverman, 2013).

Kim et al. (2009) found significant improvements in coded social behavior in MI. Simpson et al. (2015) also found that levels of engagement were significantly higher in the music condition. Katagiri (2009) found that the music conditions led to the most important improvements in emotion recognition. Srinivasan et al. (2015) found that effects were significant for positive and negative affect, whereby the latter decreased more in the MI than in the comparison group. However, interested affect remained the highest in the comparison group. And while social verbalizations increased the most in the music condition, it did not differ significantly from the comparison condition (Srinivasan et al., 2016b).

To summarize, according to the selected studies, MIs were superior to alternative conditions in eight measures, but no significant difference between conditions was found in six.

Outcomes in maladaptive behaviors

Eight studies analyzed measures of maladaptive behavior outcomes in this review's definitions. In their 2008 study, Kim et al. used the Pervasive Developmental Disorder Behavior Inventory (PDDBI; Cohen & Sudhalter, 1999), adapted for a Korean population. The social approach subscale of both the teacher and parent-version were used. Sharda et al. (2018) measured problematic autistic behavior using the maladaptive behavior subscale of the VABS (VABS-MB), which was administered to parents and caregivers. The Detrimental Social Behavior (DSB) subscale of the ASSP used by Schwartzberg and Silverman in their 2013 study was also included in the maladaptive behavior outcomes.

In terms of behavioral coding, Kim, Wigram and Gold measured instances of no response in their 2009 study when participants ignored therapeutic bids. Simpson et al. (2015) measured challenging behavior presented during their intervention. Srinivasan et al. (2015)

coded for sensory, negative and stereotyped behaviors. Sensory behaviors included any self-stimulatory behavior or inappropriate use of objects. Negative behaviors included any self-injurious behavior, moments of non-compliance, and inappropriate social behaviors. Finally, stereotyped behaviors involved repetitive self-stimulating body movements. In their first 2016 paper, Srinivasan et al. examined training-specific changes in attention patterns, differentiating whether attention was focused on the social partner, an object or elsewhere. In their second paper, the authors coded for responses to three custom-developed structured social bids per recorded session, namely a question regarding their daily routine, one about their favorite object, and a general knowledge question.

Summary of findings in maladaptive behavior outcomes

Positive changes in maladaptive behavior were noted in five measures. The VABS-MB showed a significant decrease in maladaptive behavior (Sharda et al., 2018). Kim et al. (2009) found less instances of no response in the music condition. In Srinivasan and colleagues' studies, there were decreased sensory behaviors in the MI condition, more attention to social partners, and significant increases in word count over time (2015; 2016a; 2016b). Another five measures did not find benefits for MI. Neither the PDDBI (Kim et al., 2008) nor the ASSP DSB (Schwartzberg & Silverman, 2013) identified differences between intervention conditions. Simpson et al. (2015) also did not find any differences in challenging behavior between sung or spoken condition. Srinivasan's groups found more negative behaviors in early MI sessions, but no difference in sensory behaviors between MI and the comparison intervention (2015).

Outcomes in language

Six studies looked into language outcomes in MI. Sharda et al. (2018) used the PPVT-4 to test for receptive vocabulary. To assess language skills and vocabulary comprehension and production, Thompson et al. (2013) used the MacArthur Bates Communicative Development Inventory, Words and Gestures (MBCDI-WG; Fenson, 2007). The authors acknowledged that this scale was designed for neurotypical children aged up to 18 months, but deemed it suitable for older children with severe language deficits. Lim (2010); Lim and Draper (2011) designed their own scale for assessment of language production on four speech components, called the Verbal Production Evaluation Scale (VPES).

Sandiford et al. (2012) measured increase in language outcomes through observation of verbal and imitative attempts and number of correct words uttered by the participants. Simpson et al. (2015) also measured for correct responses in their intervention and compared it to moments of engagement and challenging behavior.

Summary of findings in language outcomes

There was consensus among all studies of language skills that there was a significant increase in performance for both the music and control condition. Simpson et al. (2015) noted that there was a correlation between engaged behavior and language increase, but not with challenging behavior.

Quality assessment

Considering that systematic reviews are generally concerned with reducing and flagging bias, the present quality assessment provided support for the systematic review by informing the interpretation of quantitative results and recommendations for improvements in future research. This is done in two steps. First, risk of bias in the included studies was assessed with reference to specific elements of study design. Second, quality of interventions was assessed in order to inform about essential aspects that affect the reliability of included MIs.

Risk of bias

The risk of bias assessment showed that studies are generally efficient at confirming the diagnosis of their participants. Studies that are flagged with “unknown” risk of bias stated that they examined previous clinical records of a participant’s ASD diagnosis rather than performing any confirmatory evaluation within the study. Katagiri (2009) did not report how diagnosis was controlled for.

Reporting bias was also low in the present study sample. Katagiri (2009) was flagged because it did not report means or effect sizes. Kim et al. (2008) did not report means, but reported effect sizes. Lim and Draper (2011) described a pre-post design for their study, but only reported post-intervention results. Sandiford et al. (2012) was flagged as unknown risk of bias because post-intervention measures were reported as mean changes from baseline, rather than mean of raw measure values at post-intervention assessment.

Most studies reportedly randomized participants to groups. However, Katagiri (2009) did not mention randomization in their study procedures. All studies flagged with unknown risk of bias omitted an explanation of the means of randomizing participants.

Except for three studies (Simpson et al., 2015; Srinivasan et al., 2015; Thompson et al., 2013), all studies reportedly used blinding on assessors at least. It was assumed that Srinivasan et al. (2015) blinded their assessors, but failed to report it, since their follow-up studies all reported blinding. Ghasemtabar et al. (2015), Katagiri (2009) and Schwartzberg and Silverman (2013) did not describe who was blinded.

According to study reports, the drop-out rate was above 20% for four studies (Kim et al., 2008, 2009; Schwartzberg & Silverman, 2013; Simpson et al., 2015). Two studies did not report drop-out rates (Lim, 2010; Lim & Draper, 2011). More problematic is that 10 studies failed to report how they treated drop-outs in their analysis. This omission could potentially inhibit the reliability of reported results and is referred to as attrition bias.

The majority of studies reported a measure of symptom severity for their participants. Gattino et al. (2011) and Simpson et al. (2015) failed to report the results of this assessment. The remaining five studies did not test symptom severity of their participants. It also has to be noted that even though 14 studies measured for symptom severity, three of them (Kim et al., 2008, 2009; Lim, 2010) did not use these results in their matching process. Two studies (Ghasemtabar et al., 2015; Simpson et al., 2015) stated that symptom severity was measured, but did not report these results.

Studies received a label of low risk of bias if they matched at least on age and symptom severity. This was the case for nine studies, meaning that just over half of the studies did not match based on these criteria.

Baseline characteristics were not reported in 9 studies. This is problematic for studies assessing changes over time, as it inhibits assessment of how much the intervention changed target behavior since its beginning.

Intelligence quotient (IQ) and socio-economic status (SES) are underreported in the present study sample, with only three reports of IQ (Bieleninik et al., 2017; Crawford et al., 2017; Sharda et al., 2018) and two of SES (Bieleninik et al., 2017; Sharda et al., 2018). While it can be argued that IQ as a measure of intellectual functioning is not adequate in an autistic population and can be replaced by other measures of symptom severity, SES is crucial to determine the reach of an intervention, as often studies are biased towards higher affluent families.

The quality assessment also showed that only four studies had samples larger than 20 participants per group (Bieleninik et al., 2017; Lim & Draper, 2011; Sharda et al., 2018; Simpson et al., 2015). This cut-off, while somewhat arbitrary, was chosen in consideration of drastic reduction in statistical reliability and generalizability arising in smaller sample sizes.

Study ID	Diagnosis Control	Sample >20*	Randomization	Symptom Severity	IQ	SES	Matching**	Baseline Characteristics	Blinding	Drop-Out Rate	Attrition Bias	Reporting Bias
Bieleninik 2017	+	+	+	+	+	+	+	+	+	+	+	+
Crawford 2017	+	−	+	+	+	−	+	+	+	+	−	+
Gattino 2011	+	−	+	?	?	−	−	+	+	+	−	+
Ghasemtabar 2015	+	−	?	+	−	−	+	+	?	+	+	+
Katagiri 2009	−	−	−	−	−	−	−	−	?	+	−	−
Kim 2008	+	−	?	+	−	−	−	−	+	−	+	?
Kim 2009	+	−	?	+	−	−	−	−	+	−	−	+
Lim 2010	+	−	?	+	−	−	−	−	+	?	−	+
Lim 2011	?	+	+	−	−	−	−	−	+	?	−	?
Sandiford 2013	+	−	+	+	−	−	+	+	?	+	+	+
Schwartzberg 2013	?	−	+	−	−	−	−	+	+	−	−	+
Sharda 2018	+	+	+	+	+	+	+	+	+	+	−	+
Simpson 2013	+	+	?	?	−	−	−	−	−	−	−	+
Srinivasan 2015	+	−	?	+	−	−	+	−	−	+	+	+
Srinivasan 2016a	+	−	?	+	−	−	+	−	+	+	+	+
Srinivasan 2016b	+	−	?	+	−	−	+	−	+	+	+	+
Thompson 2013	+	−	+	+	−	−	+	+	−	+	−	+

Figure 2: Risk of bias assessment

Quality of interventions

The intervention quality assessment revealed that generally, studies fared well on explaining the rationale and supporting theoretical framework for their intervention, as well as providing an intervention protocol allowing for replication, describing adequate qualification of the person administering the intervention, adequacy of control condition and adequacy of measures used to assess intervention outcomes. Bieleninik et al. (2017) did not provide much of a theoretical framework in support of their intervention and did not provide a protocol for their intervention. Ghasemtabar et al. (2015), Katagiri (2009) and Simpson et al. (2015) did not list any details about who administered their interventions and how intervention quality was assured. Another five studies only provided partial information (Gattino et al., 2011; Kim et al., 2008, 2009; Lim, 2010; Sandiford et al., 2012) about caregiver qualification, and two about control of intervention consistency (Kim et al., 2008, 2009). Additionally, Ghasemtabar et al. (2015)'s study did not have a control condition other than a waitlisted group, which is a questionable control to assess efficacy of one specific intervention type. Studies were considered of unknown quality if they provided some information, but not enough to allow for replicability of their intervention. In the case of caregiver qualification, studies were considered of unknown quality if the interventionist was a student without specifying their years of experience. This is more problematic for those studies that do not specify if and how they controlled for consistency of the given treatment. Four studies were also flagged with unknown quality because they employed a customized measure that had not been validated previously, as this limits generalizability of their outcome data.

The two largest weaknesses of the present sample were a lack of testing for consistency of the interventions, and the short durations of all interventions. Only five studies explained how they evaluated consistency of therapy (Bieleninik et al., 2017; Crawford et al., 2017; Gattino et al., 2011; Sharda et al., 2018; Thompson et al., 2013). This was either accomplished by video-recording sessions or featuring a contingency plan for the intervention. Kim et al. (2008, 2009) had a semi-flexible treatment manual, but did not give more details for evaluation.

Most interventions lasted less than 10 hours, which inhibits the impact of treatment outcomes over time. Seven studies lasted at least 10 hours; Bieleninik et al. (2017) and Crawford et al. (2017) had the longest interventions, lasting over 32 hours, Srinivasan et al. (2015;2016a; 2016b)'s interventions lasted 24 hours, Ghasemtabar et al. (2015)'s intervention lasted 12, and Thompson et al. (2013)'s only about 10 hours total. Since these durations are still rather brief, they warranted a rating of unknown quality for these studies.

Study ID	Intervention Rationale	Intervention Protocol	Intervention Duration in *h*	Consistency of	Caregiver Qualification	Adequacy of Control	Adequacy of Measure
Bieleninik 2017	–	–	?	+	+	+	+
Crawford 2017	+	?	?	+	+	+	+
Gattino 2011	+	?	–	+	?	+	+
Ghasemtabar 2015	+	?	?	–	–	–	+
Katagiri 2009	+	?	–	–	–	+	?
Kim 2008	+	?	–	?	?	+	+
Kim 2009	+	+	–	?	?	+	+
Lim 2010	+	+	–	–	?	+	?
Lim 2011	+	+	–	–	+	+	?
Sandiford 2013	+	+	–	–	?	+	+
Schwartzberg 2013	+	?	–	–	+	+	?
Sharda 2018	+	?	–	+	+	+	+
Simpson 2013	+	?	–	–	–	+	+
Srinivasan 2015	+	+	?	–	+	+	+
Srinivasan 2016a	+	+	?	–	+	+	+
Srinivasan 2016b	+	+	?	–	+	+	+
Thompson 2013	+	+	?	+	+	+	+

Figure 3: Quality assessment

Results of meta-analyses

Selected studies

As detailed above, 13 studies met the criteria for quantitative analyses and are listed in Table 1. Outcome measures were categorized as “prosocial”, “mixed social”, “social maladaptive”, “non-social maladaptive”, or “language”. Analyses were performed individually for these measure categories, but started with the more general measure groupings of “overall”, “overall social”, “overall maladaptive”, and “language”. Analyses were performed separately for longitudinal and within-session designs, and sensitivity to extremes of multiple-measure correlation imputation assumptions was also assessed in supplemental analyses as described in the methods. The set of studies in each meta-analysis is shown in Table 3 along with a summary of the statistical results. Forest plots of the meta-analytic results are included in Figure 4 (overall across all measure categories), Figure 5 (overall social measures), Figure 6 (overall maladaptive and language measures), and the Appendix (all remaining individual measure categories).

Overall MI effect

Within session

Four studies assessed intervention outcomes in a within-session design. A total of 15 effect sizes were included from these studies, and encompassed the measure categories of prosocial, social maladaptive, non-social maladaptive, and language. The main analysis did not find a difference between MI and NMI (*SMD* = 0.13, *CI* = -0.04 to 0.31, *p* = .142; Figure 4). In a sensitivity analysis with maximal imputed r values, a significant effect size was observed (*SMD* = 0.16, *CI* = 0.04 to .27, *p* = .008).

Table 3: Results of meta-analyses. Statistics are shown for the main analyses as well as sensitivity analyses performed at minimal and maximal correlation (Pearson r) imputation levels.

Outcomes	**Design**	**Included Studies**	***n* of studies**	***n* of effect sizes**	**Imputed *r* value**	**SMD**	**95% CI**	***p***
All Outcomes	Within Session	Kim 2009; Lim 2011; Simpson 2013; Srinivasan 2015	4	15	.500 (main) .000 (min) .990 (max)	0.13 0.11 0.16	-0.04 to 0.31 -0.10 to 0.32 0.04 to 0.27	.142 .312 .008*
	Longitudinal (MI vs NMI)	Bieleninik 2017; Crawford 2017; Gattino 2011; Lim 2010; Schwartzberg 2013; Sharda 2018; Thompson 2014	7	16	.753 (main) .272 (min) .999 (max)	0.15 0.12 0.13	-0.02 to 0.32 -0.03 to 0.27 -0.01 to 0.27	.082 .106 .069
	Longitudinal (MI only)	Bieleninik 2017; Crawford 2017; Gattino 2011; Lim 2010; Schwartzberg 2013; Sharda 2018; Thompson 2014	7	16	.753 (main) .272 (min) .999 (max)	0.32 0.32 0.33	0.02 to 0.63 0.01 to 0.62 0.02 to 0.63	.038* .041* .035*
	Longitudinal (NMI only)	Bieleninik 2017; Crawford 2017; Gattino 2011; Lim 2010; Schwartzberg 2013; Sharda 2018; Thompson 2014	7	16	.753 (main) .272 (min) .999 (max)	0.09 0.08 0.09	-0.18 to 0.35 -0.18 to 0.35 -0.16 to 0.34	.522 .536 .499
All Social	Within Session	Kim 2009; Simpson 2013; Srinivasan 2015	3	11	.500 (main) .000 (min) .990 (max)	0.31 0.26 0.37	0.05 to 0.56 -0.01 to 0.53 0.11 to 0.62	.017* .062 .006*
	Longitudinal (MI vs NMI)	Bieleninik 2017; Crawford 2017; Gattino 2011; Schwartzberg 2013; Sharda 2018; Thompson 2014	6	13	.753 (main) .272 (min) .999 (max)	0.14 0.11 0.13	-0.08 to 0.35 -0.09 to 0.31 -0.06 to 0.32	.218 .267 .181
	Longitudinal (MI only)	Bieleninik 2017; Crawford 2017; Gattino 2011; Schwartzberg	6	13	.753 (main) .272 (min) .999 (max)	0.18 0.15 0.21	-0.01 to 0.36 -0.01 to 0.30 -0.02 to 0.45	.063 .061 .077

		2013; Sharda 2018; Thompson 2014						
	Longitudinal (NMI only)	Bieleninik 2017; Crawford 2017; Gattino 2011; Schwartzberg 2013; Sharda 2018; Thompson 2014	6	13	.753 (main) .272 (min) .999 (max)	0.00 -0.01 0.01	-0.17 to 0.16 -0.18 to 0.17 -0.13 to 0.15	.989 .945 .866
Pro-Social	Within Session	Kim 2009; Simpson 2013; Srinivasan 2015	3	7	.500 (main) .000 (min) .990 (max)	0.43 0.39 0.46	-0.20 to 1.05 -0.22 to 1.00 -0.17 to 1.10	.180 .210 .153
	Longitudinal (MI vs NMI)	Gattino 2011; Schwartzberg 2013; Sharda 2018; Thompson 2014	4	5	.753 (main) .272 (min) .999 (max)	0.42 0.40 0.45	-0.82 to 1.67 -0.82 to 1.62 -0.83 to 1.72	.508 .521 .493
	Longitudinal (MI only)	Gattino 2011; Schwartzberg 2013; Sharda 2018; Thompson 2014	4	5	.753 (main) .272 (min) .999 (max)	0.36 0.32 0.41	-0.52 to 1.25 -0.51 to 1.15 -0.53 to 1.35	.419 .445 .394
	Longitudinal (NMI only)	Gattino 2011; Schwartzberg 2013; Sharda 2018; Thompson 2014	4	5	.753 (main) .272 (min) .999 (max)	0.10 0.10 0.09	0.07 to 0.13 0.07 to 0.13 0.06 to 0.12	<.001* <.001* <.001*
Mixed Social	Longitudinal (MI vs NMI)	Bieleninik 2017; Crawford 2017; Sharda 2018; Thompson 2014	4	6	.753 (main) .272 (min) .999 (max)	0.08 0.08 0.15	-0.06 to 0.23 -0.07 to 0.22 -0.02 to 0.33	.263 .317 .088
	Longitudinal (MI only)	Bieleninik 2017; Crawford 2017; Sharda 2018; Thompson 2014	4	6	.753 (main) .272 (min) .999 (max)	0.16 0.15 0.19	0.06 to 0.25 0.05 to 0.25 0.09 to 0.30	.002* .003* < .001*
	Longitudinal (NMI only)	Bieleninik 2017; Crawford 2017; Sharda 2018; Thompson 2014	4	6	.753 (main) .272 (min) .999 (max)	0.03 0.02 0.03	-0.20 to 0.25 -0.22 to 0.27 -0.15 to 0.22	.822 .862 .724
Maladaptive Behaviors	Within Session	Kim 2009; Simpson 2013; Srinivasan 2015	3	6	.500 (main) .000 (min) .990 (max)	-0.06 -0.05 -0.09	-0.37 to 0.26 -0.39 to 0.29 -0.18 to -0.01	.734 .784 .024*
Social Maladaptive	Within Session	Kim 2009; Simpson 2013; Srinivasan 2015	3	4	.500 (main) .000 (min)	0.04 0.05	-0.46 to 0.54 -0.51 to 0.62	.877 .857

					.990 (max)	-0.01	-0.39 to 0.38	.968
	Longitudinal (MI vs NMI)	Schwartzberg 2013; Sharda 2018	2	2	.753 (main) .272 (min) .999 (max)	-0.05 -0.05 0.07	-0.14 to 0.05 -0.15 to 0.04 -0.17 to 0.31	.312 .267 .580
	Longitudinal (MI only)	Schwartzberg 2013; Sharda 2018	2	2	.753 (main) .272 (min) .999 (max)	0.23 0.23 0.19	0.16 to 0.30 0.16 to 0.31 0.11 to 0.27	<.001* <.001* <.001*
	Longitudinal (NMI only)	Schwartzberg 2013; Sharda 2018	2	2	.753 (main) .272 (min) .999 (max)	0.09 0.18 0.05	-0.30 to 0.49 -0.14 to 0.50 -0.35 to 0.45	.642 .271 .806
Language	Within Session	Lim 2011; Simpson 2013	2	2	.500 (main) .000 (min) .990 (max)	0.08 0.09 0.00	-0.04 to 0.21 -0.04 to 0.21 -0.18 to 0.18	.195 .165 .988
	Longitudinal (MI vs NMI)	Lim 2010; Sharda, 2018; Thompson, 2014	3	3	.753 (main) .272 (min) .999 (max)	0.02 0.00 0.16	-0.09 to 0.13 -0.03 to 0.03 -0.08 to 0.39	.727 .969 .188
	Longitudinal (MI only)	Lim 2010; Sharda, 2018; Thompson, 2014	3	3	.753 (main) .272 (min) .999 (max)	0.59 0.58 0.60	-0.10 to 1.28 -0.15 to 1.32 -0.06 to 1.26	.092 .120 .075
	Longitudinal (NMI only)	Lim 2010; Sharda, 2018; Thompson, 2014	3	3	.753 (main) .272 (min) .999 (max)	0.43 0.42 0.44	-0.11 to 0.97 -0.17 to 1.01 -0.06 to 0.94	.118 .159 .086

Longitudinal

Seven studies measured outcomes in a longitudinal design. A total of 16 effect sizes were included from these studies, and encompassed the measures of prosocial, mixed social, social maladaptive, and language. There was a non-significant trend toward greater benefits in MI than NMI (*SMD* = 0.15, *CI* = -0.02 to 0.32, *p* = .082; Figure 4). Change between pre-intervention and post-intervention measurements was significant within music interventions (*SMD* = 0.32, *CI* = 0.02 to 0.63, *p* = .038; Figure 4), but not in control interventions (*SMD* = 0.09, *CI* = -0.18 to .35, *p* = .522; Figure 4).

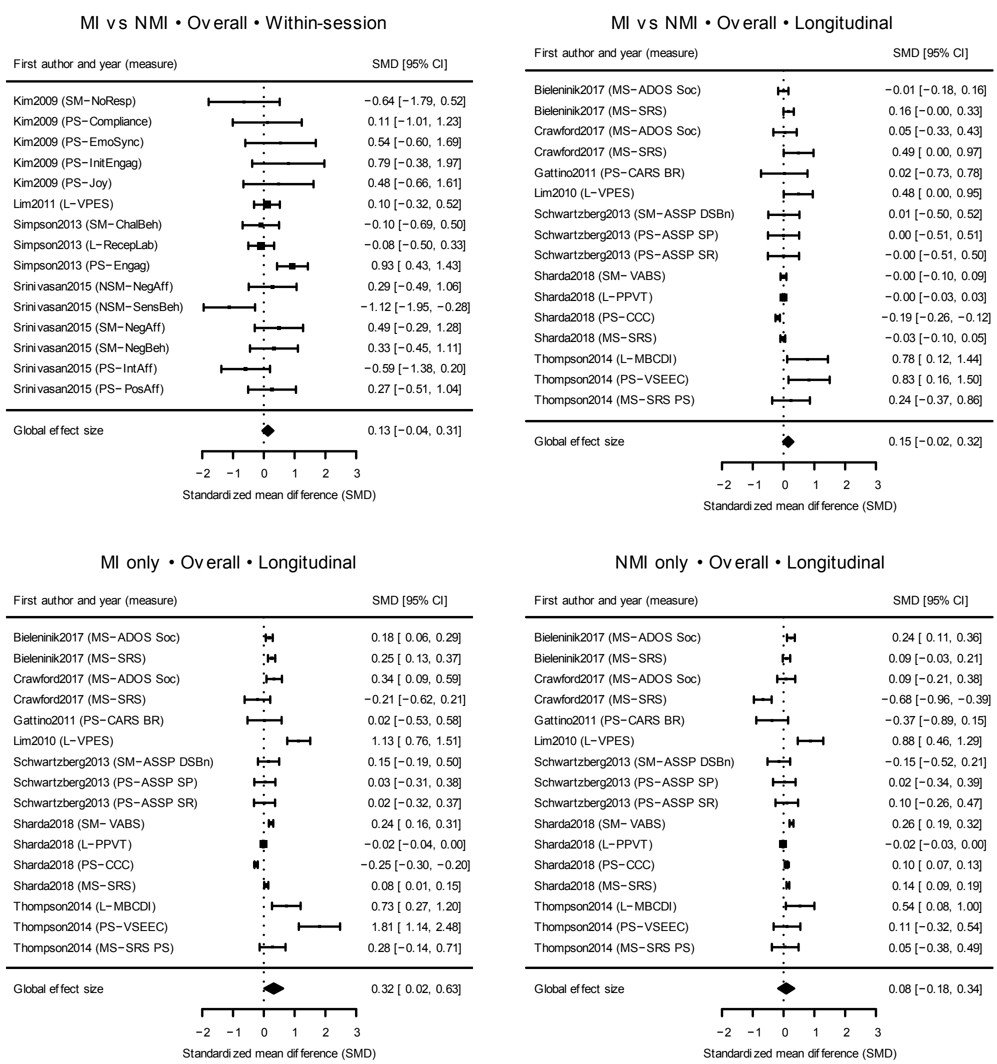


Figure 4: Forest plots of analysis of all study outcomes

Effects of MI on social outcomes

Within-session

Three studies used a within-session design to measure social outcomes from MI and NMI. There was a significant effect of intervention across all within-session social outcomes (*SMD* = 0.31, *CI* = 0.05 to 0.56, p = .017; Figure 5), except in a sensitivity analysis of minimal imputed r values, which showed a non-significant trend (*SMD* = 0.26, *CI* = -0.01 to 0.56, p = .062). No significant effect was found for prosocial measures on three studies (*SMD* = 0.43, *CI* = -0.20 to 1.05, p = .180).

Longitudinal

Seven studies assessed longitudinal effects of MI on social outcomes. Across all social outcomes, there was no difference in intervention effect between MI and NMI (*SMD* = 0.14, *CI* = -0.08 to 0.35, p = .218; Figure 5). A non-significant trend was observed within music interventions (*SMD* = 0.18, *CI* = -0.01 to 0.36, p = .063; Figure 5), but not within alternative interventions (*SMD* = 0.00, *CI* = -0.17 to 0.16, p = .989; Figure 5).

Five longitudinal studies examined prosocial behavior changes. There was no difference in prosocial intervention effect between MI and NMI (*SMD* = 0.42, *CI* = -0.82 to 1.67, p = .508), nor within MI alone (*SMD* = 0.36, *CI* = -0.52 to 1.25, p = .419), but prosocial measures did improve in NMI (*SMD* = 0.10, *CI* = 0.07 to 0.13, $p < .001$).

Four studies assessed changes in social behavior using mixed outcome measures such as the SRS that assessed both pro- and antisocial behaviors. There was no difference in intervention effect between MI and NMI (*SMD* = 0.08, *CI* = -0.06 to 0.23, p = .263), but benefits were found within MI (*SMD* = 0.16, *CI* = 0.06 to 0.25, p =.002) and not in (*SMD* = 0.03, *CI* = -0.20 to 0.25, p = .822).

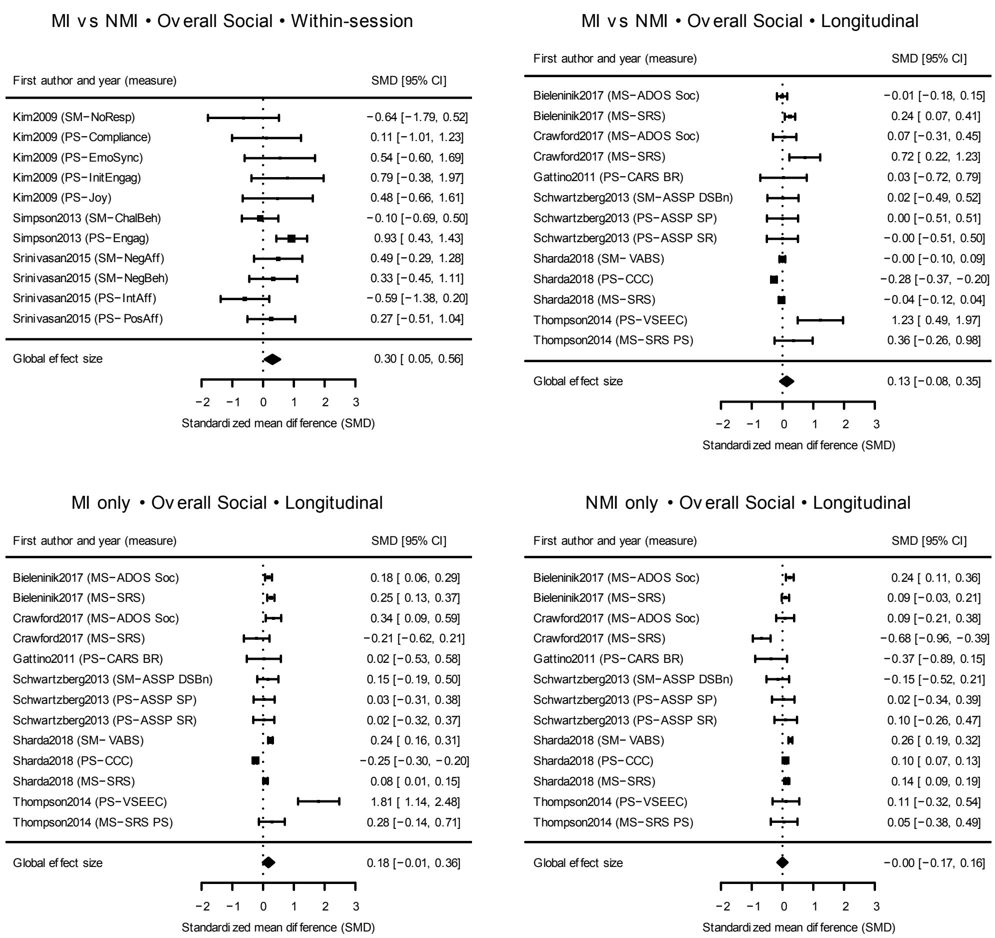


Figure 5: Forest plots of analysis of all social outcomes

Effects of MI on maladaptive behavior outcomes

Within session

Three studies examined maladaptive behavior using within-session designs. Across both social and non-social maladaptive behavior, no difference in intervention effect between MI and NMI was found (*SMD* = -0.06, *CI* = -0.37 to 0.26, *p* = .734; Figure 6), except in a sensitivity analysis of maximal imputed *r* values (*SMD* = -0.09, *CI* = -0.18 to -0.01, *p* = .024) in the direction of an advantage for NMI over MI.

Three studies investigated socially maladaptive outcomes in within-session designs. No difference in intervention effect between MI and NMI was found (*SMD* = 0.04, *CI* = -0.46 to 0.54, *p* = .877).

Longitudinal

Only two studies assessed maladaptive behavior outcomes in a longitudinal design. All of these outcomes examined socially maladaptive behavior. There was no difference in intervention effect between MI and NMI (*SMD* = -0.05, *CI* = -0.14 to 0.05, *p* = .312). However, benefits were found within MI (*SMD* = 0.23, *CI* = 0.16 to 0.30, *p* < .001), but were not significant for NMI (*SMD* = 0.09, *CI* = -0.30 to 0.49, *p* = .642).

Effects of MI on language outcomes

Within session

Three studies examined language outcomes in within-session designs. No difference in intervention effect between MI and NMI was found (*SMD* = 0.08, *CI* = -0.04 to 0.21, *p* = .195; Figure 6).

Longitudinal

Three studies assessed language outcomes in a longitudinal design. No difference was found between MI and NMI (*SMD* = 0.02, *CI* = -0.09 to 0.13, *p* = .727; Figure 6), nor were there significant effects for MI (*SMD* = 0.59, *CI* = -0.10 to 1.28, *p* = .092) or NMI (*SMD* = 0.43, *CI* = -0.11 to 0.97, *p* = .118).

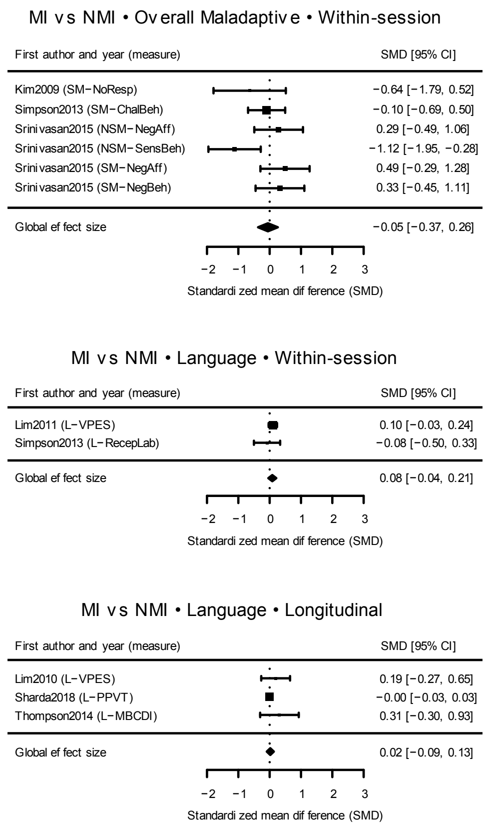


Figure 6: Forest plots of analyses on maladaptive behavior and language outcomes

Discussion

This review investigated the efficacy of MI compared to NMI on social functioning, maladaptive behavior and language outcomes in children with ASD. The goal was to investigate the hypothesis set by some studies that MIs pose a benefit over non-music interventions in this population. This question was examined in two parts: first, a systematic review of the included studies in terms of their design, methodology, reported measures and findings; and second, a set of quantitative meta-analyses to assess the efficacy of MI within key categories of social, maladaptive and language outcome.

The results are suggestive of a benefit of MI over NMI in ASD for social outcomes, but do not find benefits for non-social maladaptive behaviors or language outcomes. In the systematic review, a majority of studies reported benefits in MI compared to NMI for both social function and maladaptive behavior. For language outcomes, only a minority of studies reported benefits in MI. However, several methodological problems were common across the study sample: small sample sizes, short intervention durations, underreporting of values important for group matching, and lack of consideration of potential biases arising from attrition. Notwithstanding these issues, the quantitative meta-analyses identified several informative patterns of results that support benefits of MI over NMI. The analyses combining social, maladaptive and language outcomes resulted in near-significant effects in longitudinal studies, whereas studies of social outcomes using a within-session approach showed a significant benefit of MI over NMI. Additional support for a benefit of MI was found in the individual analyses of longitudinal effects in MI and NMI, in which MI had significant pre-post improvement but NMI was non-significant; this pattern was found for the overall analysis across all outcomes, and for the analyses of mixed social outcomes and social maladaptive behaviors, but did not hold for prosocial social outcomes.

These findings, while supportive of a possible benefit of MI over NMI, reveal methodological issues present in studies of MI in ASD children. By means of systematic review and meta-analysis, this study aims to inform the methodology of future research to provide more confident assessments of music-based interventions in ASD.

MI effects on social, maladaptive and language outcomes

The literature search identified 19 studies that met the inclusion criteria for systematic review. Twelve studies used a longitudinal design in which outcomes were measured before and after the intervention period, whereas the remaining seven employed a within-session design in which outcomes were measured during the intervention itself. The questions answered by these two designs have different nuances: comparing pre-versus-post measurements seeks to answer if a given intervention can generalize outcomes outside the therapeutic setting, whereas a within-session comparison evaluates which intervention leads to more desirable behavior during treatment. Eleven studies were included in the quantitative section of this review, of which seven adhered to a longitudinal design, and four assessed MI effects within sessions.

The narrative part of this study found that a small majority of studies found a benefit of MI over alternative interventions, or at least very similar results in both. It is interesting to note that there was no ambiguity within any particular measure: the SRS and the ADOS social score were each used in multiple studies, and their results were consistently non-significant for the comparison between MI and NMI. However, the SRS asks respondents to report behavior that occurred over the past six months (Constantino & Gruber, 2012). No study assessed in this review lasted six months or more, suggesting that the SRS may have limited sensitivity for interventions on this time scale. Additionally, it has to be taken into account that the ADOS and CARS are primarily designed as diagnostic tools. Their use for treatment outcome is therefore debated (Fletcher-Watson & McConachie, 2017; Green et al., 2010).

There were several nuances that emerged in the meta-analyses of social behavior. In the analyses of longitudinal studies, direct comparisons between MI and NMI did not clearly show a superiority for MI, but in separate analyses MI showed several key longitudinal improvements whereas NMI did not produce outcome changes over the course of the interventions except for prosocial measures. Aside from the latter result for prosocial measures, the apparent low effectiveness of NMIs in the analyzed studies was surprising because standard-of-care interventions for ASD children are well researched and are known for their positive effects on social behaviors in ASD (e.g. Dawson et al., 2010; Pickles et al., 2016; Rogers et al., 2019; Schröder et al., 2015; Warren et al., 2011; White, Keonig, & Scahill, 2007). The general lack of longitudinal effects during these standard NMI interventions in the present study sample may reflect shortcomings in the duration or implementation rather than a true lack of efficacy for these types of interventions.

The distinction in longitudinal results between prosocial measures and the other social measure categories suggests that MI may not hold particular benefits (or may in fact be less effective than NMI) for strictly prosocial behaviors, such as engagement and responding to prompts. This is reinforced by the analysis of within-session prosocial outcomes, which did not find a difference between MI and NMI. In total, there is more support in the present results for longitudinal benefits of MI on social maladaptive behaviors (as reflected in the "social maladaptive" and "mixed social" categories) than on prosocial behaviors. It is possible that MI are generally better at addressing some social behaviors whereas other social behaviors are better promoted by NMI. These results point towards the importance of isolating particular aspects of MI to discern which elements of interventions will best address particular behavioral impairments or strengths, with consideration of wide individual differences in the autism spectrum.

Looking at maladaptive behavior outcomes, some research points towards music having a soothing effect for ASD children, with the consequence of reducing maladaptive behaviors (Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004). Half of the studies in the systematic review found a benefit for MI on maladaptive behaviors, whereas the other half did not find an advantage for MI. In the meta-analyses, the question of MI's effect on overall maladaptive behaviors (including both social and non-social aspects) was only addressed by within-session studies, which did not find a difference between MI and NMI on this outcome (n.b. during sensitivity analyses, one result found greater maladaptive behaviors in MI compared to NMI). For maladaptive behaviors with specific impact on social situation outcomes, such as responsiveness to social requests, there was no significant difference between MI and NMI for within-session studies, although as noted in the previous section, longitudinal studies showed an effectiveness for MI on social maladaptive behaviors that was not found in NMI. Unfortunately, the question of MI's impact on non-social maladaptive behaviors (such as stimming) cannot be addressed in meta-analysis because only one study reported such measures.

There were no conclusive results for language outcomes. Meta-analysis did not find any significant benefit of MI over NMI for either within-session or longitudinal studies. However, both these analyses contained only three studies. Overall, it appeared that MI and NMI did not differ in effectiveness in this domain. However, participants were generally not novices to language learning interventions, and therefore carry-over effects could have influenced results. It also has to be mentioned that language outcomes in this review measured either accuracy or frequencies of verbalizations. These measure types were grouped together as a practical consequence of the limited number of studies available, but strictly speaking, these measures answer two different questions: whether MI are more effective than NMI in facilitating language training, and whether MI are more likely than NMI to elicit verbal communication in ASD children. It is possible that MI are more effective in one of these domains than the other, but more studies are needed to draw any conclusions.

Quality assessment of MI studies

The present findings were possibly subject to a number of biases in the included studies that varied strongly in key aspects such as sample sizes, sample description, matching criteria and quality and reporting of attrition. Due to different sample sizes within these studies, as well as different measures used, some studies have greater weighting than others, which likely has an influence on the present results, despite best efforts to reduce bias. The inconclusivity of results is increased by the short duration and small intensity of the analyzed studies. However, typical interventions in ASD children average 30 hours a week, last at least one year and yield significant results (e.g. Dawson et al., 2010; Rogers et al., 2019). In order to reach more definite results, studies need to be longer and more intense in both MI and NMI conditions.

Six studies failed to both report baseline characteristics and match according to at least two criteria, which resulted in omission of important sample information. Underreporting of baseline characteristics and symptom severity is problematic for longitudinal studies, as it hinders the understanding of changes in outcome measures over the full course of the intervention. Even though ASD symptom severity was measured by the majority of studies, it was often not used as a matching criterion. In fact, matching was mostly solely done on age, which is not sufficient as a criterion, as chronological age is not a reliable indicator of mental age and cognitive and social functions (Jarrold & Brock, 2004). Bias can occur when intervention group assignment is confounded with characteristics including IQ, symptom severity and age. IQ is important because it varies widely across the autism spectrum and can affect the ability to receive specific interventions (Matson, Dempsey, LoVullo, & Wilkins, 2008). Symptom severity is important because it informs the degree of impairment of various ASD markers, such as maladaptive behavior (Jang, Dixon, Tarbox, & Granpeesheh, 2011). Age is important because it serves as an indicator for expected milestones according to symptom severity and IQ (Kover, Edmunds, & Weismer, 2016). For these reasons, it is problematic when studies do not report the baseline values of these characteristics, and when they do not ensure that these values are matched between groups. A lack of transparency in matching, or matching on an insufficient set of variables, increases the apparent risk of bias as it is not possible to exclude the possibility that group differences arise from inherent differences within the sample as opposed to treatment effects (Anderson et al., 1980). Participant attrition is a common issue for longitudinal studies, and presents challenges both for assessing bias risk and for reporting data for meta-analysis. Attrition may be influenced by potentially confounding factors, such as age, cognitive level or SES, and could thereby result in poorer matching of groups at the end of a study than at baseline. No studies in this review made any explicit consideration of the demographics of individuals who did not complete the intervention, or addressed this potential source of bias. Additionally, meta-analysis of longitudinal data assumes that the pre- and post-intervention measurements are presented for the same sample of participants, but many studies in this review did not report the required baseline summary values (e.g., mean and standard deviation) for the exact set of participants who had post-intervention measures. The actual sample is thus not portrayed correctly, undermining generalizability of the outcome. The potential risk arising from this problem is exacerbated considering the study samples were generally quite small, and presents a methodological limitation on any review of these studies.

Future Directions

In consideration of the limitations identified above, future studies should aim for higher intensities and longer durations in order to generate more realistic and generalizable results. Additionally, studies should attempt to recruit larger samples and collect SES data from participating families. It is also important to match intervention groups on symptom severity and IQ in order to avoid potential confounds between intervention type and these parameters in the results. More rigorous baseline reporting, such as descriptions of symptom severity in each intervention group, is also needed. Another important addition is information about individuals who failed to complete the study, and consideration of potential biases or confounds resulting from attrition.

Future studies should consider additional fundamental questions about MI in ASD children. For instance, while the present results generally support the idea that MI are as effective as NMI, and may have benefits above NMI, there is not enough information to dissociate which elements of music may be beneficial for specific outcomes. Some studies suggest that music can help ASD children control their moods and hence modulate their behavior (Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004), but the present study did not find a significant difference between MI and NMI in terms of maladaptive behavior moderation. In fact, in one study, participants in the MI condition at least initially struggled with higher maladaptive behavior than NMI (Srinivasan et al., 2015). The lack of longitudinal investigations of maladaptive behaviors however does not allow for a definite conclusion. It is possible that other types of MI beyond the active music interventions reviewed here, such as AIT, could have such an effect (Dimitriadis & Smeijsters, 2011; Wan, Demaine, Zipse, Norton, & Schlaug, 2010; Whipple, 2004). It is therefore reasonable to expect that different uses of music in the ASD population may lead to different outcomes. It is also possible that the strongest advantages of MI over NMI reside in factors other than those measured in this study. It has for instance been suggested that MIs create more joy and life satisfaction in ASD participants than alternative interventions (Bieleninik et al., 2017; Eren, 2015; Kim et al., 2009).

Conclusions

The present systematic review and meta-analyses generally support the effectiveness of MI, and provide evidence of benefits from MI as compared to NMI for social outcomes. However, methodological issues in the currently available literature impact the conclusiveness and generalizability of many study results. Of particular note, study sample sizes were small, and durations and intensities of interventions were considerably shorter than real-life scenarios. These results provide an up-to-date evaluation of the evidence for MI's benefits in children with ASD and identify recommended best practices to strengthen future clinical research in this domain.

Discussion

This review investigated the efficacy of MI compared to NMI on social functioning, maladaptive behavior and language outcomes in ASD children. The goal was to investigate the hypothesis set by some studies that MIs pose a benefit over non-music interventions in this population. This question was examined in two parts: first, a systematic review of the included studies in terms of their design, methodology, reported measures and findings; and second, a set of quantitative meta-analyses to assess the efficacy of MI within key categories of social, maladaptive and language outcome.

The results are suggestive of a benefit of MI over NMI in ASD for social outcomes, but do not find benefits for non-social maladaptive behaviors or language outcomes. In the systematic review, a majority of studies reported benefits in MI compared to NMI for both social function and maladaptive behavior. For language outcomes, only a minority of studies reported benefits in MI. However, several methodological problems were common across the study sample: small sample sizes, short intervention durations, underreporting of values important for group matching, and lack of consideration of potential biases arising from attrition. The quantitative meta-analyses identified mainly trends toward a benefit of MI over NMI. The analyses combining social, maladaptive and language outcomes resulted in near-significant effects in longitudinal studies, whereas studies of social outcomes using a within-session approach showed a significant benefit of MI over NMI. Additional support for a benefit of MI was found in the individual analyses of longitudinal effects in MI and NMI, in which MI had significant pre-post improvement but NMI was non-significant; this pattern was found for the overall analysis across all outcomes, and for the analyses of mixed social outcomes and social maladaptive behaviors, but did not hold for prosocial social outcomes.

These findings, while suggestive of a benefit of MI over NMI, reveal methodological issues present in studies of MI in ASD children. By means of systematic review and meta-analysis, this study aims to inform the methodology of future research to provide more confident assessments of music-based interventions in ASD.

Discussion – Systematic Review

The literature search identified 19 studies that met the inclusion criteria, and these studies subsequently underwent systematic review. Twelve studies used a longitudinal design in which outcomes were measured before and after the intervention period, whereas the remaining seven employed a within-session design in which outcomes were measured during the intervention itself. The questions answered by these two designs have different nuances: comparing pre-versus-post measurements seeks to answer if a given intervention can generalize outcomes outside the therapeutic setting, whereas a within-session comparison evaluates which intervention leads to more desirable behavior during treatment. Both types of studies were analyzed here, as long as they examined at least one of the three targeted outcomes types of social function, maladaptive behavior and language.

Outcomes on social behavior

Fourteen studies compared social behavior in MI and NMI using a total of 19 different measures. Eight studies found favorable outcomes for music compared to NMI. Six studies did not find a significant benefit of MI over alternative conditions. No studies found lower social outcomes in MI compared to NMI.

It is interesting to note that there was no ambiguity within any particular measure: the SRS and the ADOS social score were each used in multiple studies, and their results were consistently non-significant. However, the SRS asks respondents to report behavior that occurred over the past six months (Constantino & Gruber, 2012). No study assessed in this review lasted six months or more, rendering the sensitivity of SRS results questionable. Additionally, it has to be taken into account that the ADOS and CARS are primarily designed as diagnostic tools. Their use for treatment outcome is therefore debated (Fletcher-Watson & McConachie, 2017; Green et al., 2010). Another factor to consider is heterogeneity across the social measures used in the study sample. Nine different standardized measures and 10 different observation codings were used across 14 studies. The fact that there is arguably some variability across those measures could potentially explain the differences in results, as could variability in characteristics of interventions and participant samples (e.g., intervention nature, duration, intensity and context, and participant age and symptom severity).

It can still be argued that a small majority of measures found significant effects in favor of MI, therefore replicating findings of Accordino et al. (2007) and Geretsegger et al. (2014). Even though some studies did not find MI to be superior to standard care, MI often yielded similar results (Gattino et al., 2011; Schwartzberg & Silverman, 2013; Sharda et al., 2018; Srinivasan et al., 2016b), which supports their value as alternative interventions.

Outcomes on maladaptive behavior

Eight studies measured maladaptive behavior in a total of 10 different measures. Half found a significant benefit for MI and the other half did not; no studies found lower maladaptive outcomes in MI compared to NMI. It is important to note that most studies looked at maladaptive behavior in terms of behavior within the sessions. From these data, it is therefore not possible to draw conclusions as to whether or not MI could lead to enduring behavioral changes outside this setting, a limitation previous raised by Accordino et al. (2007).

Another factor to take into consideration is that the MIs analyzed here used music in an active manner, requiring participants to interact with the musical output or create it themselves. It is possible that greater effects on behavior modulation could be found when music is used as a background stimulus, a usage of music that was not in the scope of the present review. Indeed, autistic people often report that they utilize background music to moderate their emotional states (Allen et al., 2009) and studies using modulated forms of background music (e.g., AIT) have reported improvements in agitation, anxiety and attention (Dimitriadis & Smeijsters, 2011; Wan et al., 2010; Whipple, 2004). It is therefore reasonable to assume that not all music or all kinds of interactions with music may lead to a reduction in maladaptive behavior. It is clear that more longitudinal studies on MI effects on both social and non-social maladaptive behavior in ASD are needed.

An additional consideration about non-standard interventions such MI is a potential need for participants to acclimatize to its novel aspects. Srinivasan et al. (2015) initially measured greater maladaptive behavior in early sessions of their MI compared to the control condition. Over time, these behaviors decreased more significantly in MI than in their robot and control condition, but on an absolute basis they did not drop below levels measured in the control (seatwork) condition. The authors reasoned that initially high levels of maladaptive behaviors observed in the music condition could simply be due to the novelty of the intervention. This aspect is important to consider both for the implementation of a MI, and in understanding differences in within-session behavioral measures between MI and NMI when the study duration may not provide sufficient time for acclimatization to the novelty of an unfamiliar intervention.

Outcomes on language

Six types of measures were used to evaluate language outcomes. In all six studies examining language, there was no difference found between MI and NMI, although in the four studies of longitudinal change there was a significant increase in language measures in both music and control conditions. The result that MIs have no benefits over NMIs on language and verbalizations seems counter-intuitive, considering the links drawn between music and language by some studies (e.g. Accordino et al., 2007; Buday, 1995; Geretsegger et al., 2014), as well as the facilitating effects of intonation, rhythmicality and pitch upon language learning (LaGasse, 2014; Lim, 2010). However, it has to be noted that participants in the present study sample were typically not novices to standard care practice, and the lack of difference in outcomes between both types of interventions could be due to larger practice effects in favor of the non-music intervention. This potential confound makes it more difficult to draw conclusions. It is also possible that more clear differences in language outcome between MI and NMI would become apparent if studies followed longer periods of intervention.

Risk of bias

As shown in the risk of bias assessment (Figure 2), eight of the 17 studies present a moderate to high risk of bias on 6 or more criteria. Most problematic were the small sample sizes. Additional risks included underreporting of IQ and SES, matching on only one criterion, omitting outcomes at baseline, and not reporting how attrition biases were considered.

Even though ASD symptom severity was measured by the majority of studies, it was often not used as a matching criterion. In fact, matching was mostly solely done on age, which is not sufficient as a criterion, as chronological age is not a reliable indicator of mental age and cognitive and social functions (Jarrold & Brock, 2004). Bias can occur when intervention group assignment is confounded with characteristics including IQ, symptom severity and age. IQ is important because it varies widely across the autism spectrum and can affect the ability to receive specific interventions (Matson et al., 2008). Symptom severity is important because it informs the degree of impairment of various ASD markers, such as maladaptive behavior ((Jang et al., 2011). Age is important because it serves as an indicator for expected milestones according to symptom severity and IQ (Kover et al., 2016). For these reasons, it is problematic when studies do not report the baseline values of these characteristics, and when they do not ensure that these values are matched between groups. Six studies failed to both report baseline characteristics and match according to at least two criteria, which resulted in omission of important sample information. Underreporting of baseline characteristics and symptom severity is problematic for longitudinal studies, as it hinders the understanding of changes in outcome measures over the full course of the intervention. A lack of transparency in matching, or matching on an insufficient set of variables, increases the apparent risk of bias as it is not possible to exclude the possibility that group differences arise from inherent differences within the sample as opposed to treatment effects (Anderson et al., 1980).

Participant attrition is a common issue for longitudinal studies, and presents challenges both for assessing bias risk and for reporting data for meta-analysis. Attrition may be influenced by potentially confounding factors, such as age, cognitive level or SES, and could thereby result in poorer matching of groups at the end of a study than at baseline. No studies in this review made any explicit consideration of the demographics of individuals who did not complete the intervention, or addressed this potential source of bias. Additionally, meta-analysis of longitudinal data assumes that the pre- and post-intervention measurements are presented for the same sample of participants, but many studies in this review did not report the required baseline summary values (e.g., mean and standard deviation) for the exact set of participants who had post-intervention measures. The actual sample is thus not portrayed correctly, undermining generalizability of the outcome. The potential risk arising from this problem is exacerbated considering the study samples were generally quite small, and presents a methodological limitation on any review of these studies.

Quality of interventions

Studies were graded according to number of positive, low-risk ratings. Studies having seven or more positive ratings were considered of satisfactory quality, while studies having 9 or more positive ratings were considered of good quality. Subsequently, any study scoring less than seven positive ratings were considered problematic. According to this rating-scheme, eleven of the 17 studies were scored as having satisfactory to good intervention quality, whereas 6 were of low quality. Interventions were considered problematic if they had three or more outcomes of low to unknown quality.

Caregiver qualification was a particular issue in the present study sample. Students were often charged with administering the intervention, but their years of training or practice were not reported. Another key problem was failure to describe how a study ensured adherence to the intervention protocol. If caregivers do not consistently adhere to the intervention protocol, it is impossible to interpret intervention outcomes. Intervention protocols themselves were also frequently underreported. Studies only generally explained the nature of their interventions without providing enough information to allow for reproducibility. This underreporting inhibits the assessment of intervention components responsible for successful outcomes.

The biggest drawback of the present study sample, however, was the limited intervention durations. Standard interventions in children with ASD are widely studied and have considerable evidence for their efficacy (e.g. Dawson et al., 2010; Pickles et al., 2016; Rogers et al., 2019; Schröder et al., 2015; Warren et al., 2011; White et al., 2007). However, these interventions typically take 20 to 40 hours a week and last at least two years (Vismara & Rogers, 2010). It is reasonable to assume that to achieve a similar efficacy, a comparable intervention intensity would be necessary for MI in ASD. It is likely that the findings of present MI studies are not conclusive because neither the MI nor control intervention were long and intense enough to be representative of interventions as they are conducted in clinical practice, and that studies finished before substantial changes in behavioral outcome could be measured.

Discussion – Meta-Analyses

The quantitative part of this review assessed the efficacy of MI on social outcomes, maladaptive behaviors and language in ASD children as compared to NMI. As reported earlier, analyses were conducted separately for within-session studies and longitudinal studies, because outcomes were tested differently in these designs. The meta-analyses started by assessing all intervention outcomes in within-session and pre-post interventions respectively, and then tested the individual outcome types of social functioning, maladaptive behavior and language.

A total of 11 studies were included in the quantitative section of this review. Generally, the results support the interpretation that there is some benefit of MI over NMI. In within-session studies, MI were significantly superior to NMI in general social behavior outcomes, but not on strictly prosocial outcomes. There was no significant result for within-session MI outcomes in maladaptive behaviors, nor in language outcomes. However, these analyses were small, with only three to four included studies. In longitudinal studies, direct comparison of MI to NMI did not yield significant results, but there was a pattern that MI had significant improvements in outcome, whereas NMI did not show improvements, when both types of interventions were analyzed individually. The main exception to this pattern was found in prosocial outcomes, where NMI showed longitudinal improvements but MI did not. In the following section, these results will be discussed for each outcome type.

Outcomes on social behavior

Results generally pointed toward a benefit of MI on social behavior outcomes in ASD children. However, there are several nuances in the social behavior results. In the analyses of longitudinal studies, direct comparisons between MI and NMI did not clearly show a superiority for MI, but in separate analyses MI showed several key longitudinal improvements whereas NMI did not produce outcome changes over the course of the interventions except for prosocial measures. Aside from the latter result for prosocial measures, the apparent low effectiveness of NMIs in the analyzed studies was surprising because standard-of-care interventions for ASD children are well researched and are known for their positive effects on social behaviors in ASD (e.g. Dawson et al., 2010; Rogers et al., 2019; Schröder et al., 2015; Warren et al., 2011; White et al., 2007). The general lack of longitudinal effects during these standard NMI interventions in the present study sample may reflect shortcomings in the duration or implementation rather than a true lack of efficacy for these types of interventions.

The distinction in longitudinal results between prosocial measures and the other social measure categories suggests that MI may not hold particular benefits (or may in fact be less effective than NMI) for strictly prosocial behaviors, such as engagement and responding to prompts. This is reinforced by the analysis of within-session prosocial outcomes, which did not find a difference between MI and NMI. In total, there is more support in the present results for longitudinal benefits of MI on social maladaptive behaviors (as reflected in the "social maladaptive" and "mixed social" categories) than on prosocial behaviors. It is possible that MI are generally better at addressing some social behaviors whereas other social behaviors are better promoted by NMI. These results point towards the importance of isolating particular aspects of MI to discern which elements of interventions will best address particular behavioral impairments or strengths, with consideration of wide individual differences in the autism spectrum.

Outcomes on maladaptive behavior

Some research points towards music having a soothing effect on ASD children, reducing maladaptive behaviors (Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004). However, the question of MI's effect on overall maladaptive behaviors (including both social and non-social aspects) was mostly addressed by within-session studies, which did not find a difference between MI and NMI on this outcome (n.b. during sensitivity analyses, one result found greater maladaptive behaviors in MI compared to NMI). For maladaptive behaviors with bigger impact on social situation outcomes, such as responsiveness to social requests, there was no significant difference between MI and NMI for within-session studies, although as noted in the previous section, longitudinal studies showed an effectiveness for MI on social maladaptive behaviors that was not found in NMI. Unfortunately, the question of MI's impact on non-social maladaptive behaviors (such as stimming) cannot be addressed because only one study reported such measures.

Outcomes on language

There was no significant benefit of MI over NMI in language outcomes in an analysis of three within-session studies. An analysis of three longitudinal studies of language outcomes also failed to find a difference between MI and NMI. However, the small number of studies in these analyses precludes drawing any strong conclusions about the degree of benefit on language acquisition and enhancement from MI compared to NMI from these results.

General Discussion

Considering the qualitative and quantitative results together, while the narrative part of this study found that a small majority of studies found a benefit of MI over alternative interventions, or at least very similar results in both, the meta-analyses were not as conclusive. Longitudinal analyses of social measures that assessed prosocial and social maladaptive behaviors in combination (e.g., SRS), as well as analyses of social maladaptive measures alone, and analyses that included all available social measures, found that MI produced improvements whereas change during NMI was non-significant. These results support the idea that MI are as effective, or more effective than NMI (although the direct comparisons of MI to NMI in longitudinal studies did not find differences). However, this pattern was reversed in the analysis of prosocial measures alone, where NMI produced longitudinal improvements but MI had no significant change. It is unclear if this result is due to the measures used or whether MI do indeed address some social skills better than others.

Looking at maladaptive behavior outcomes, half of the studies in the systematic review found a benefit for MI, whereas the other half did not find an advantage for MI. The meta-analyses did not find benefits for MI over NMI for maladaptive behaviors. However, as

Srinivasan et al. (2015) noted, it is possible that ASD children need a longer time to acclimatize to MIs, which are less familiar to them than standard interventions. Indeed, behavioral outcomes were mostly assessed in within-session measures, which doesn't allow to draw conclusions for long-term outcomes.

There were no conclusive results for language outcomes. It appeared that MI and NMI did not differ significantly in effectiveness. However, participants were generally not novices to language learning interventions, and therefore carry-over effects could have influenced results. It also has to be mentioned that language outcomes in this review measured either accuracy or frequencies of verbalizations. These measure types were grouped together as practical consequence of the limited number of studies available, but strictly speaking, these measures answer two different questions: whether MI are more effective than NMI in facilitating language training, and whether MI are more likely than NMI to elicit verbal communication in ASD children. It is possible that MI are more effective in one of these domains than the other, but more studies are needed to draw any conclusions.

The lack of conclusiveness from the present quantitative analyses might appear unsatisfactory. However, the study limitations detailed below may serve as guidance for future studies in this domain and therefore help in answering crucial questions about MI in ASD children.

Study Limitations

The greatest limitation of this study is the small sample size. In particular, the generalizability of the meta-analyses is restricted by the small number of studies. Additionally, due to different sample sizes within these studies, as well as different measures used, some studies have greater weighting than others, which likely has an influence on the present results despite best efforts to reduce bias. The inconclusivity of the quantitative results is increased by the short duration and small intensity of analyzed studies. However, typical interventions in ASD children average 30 hours a week, last years and yield significant results. In order to reach more definite results, studies need to be longer and more intense in both MI and NMI conditions.

Another limitation is the varying methodological quality of included studies. As noted by Higgins and Green (2011), the strength of any systematic review and meta-analysis is strongly dependent on the quality of studies that are reported. Study quality varied strongly in key aspects such as sample sizes, sample description, matching criteria and quality and reporting of attrition, as reported in this review's quality assessment. Heterogeneity within interventions and outcome measures was also rather high. For instance, Bieleninik et al. (2017) studied MI effects in an active-production intervention lasting over 30 hours on over 300 participants, whereas Lim (2010) conducted a short active-listening intervention with 36 participants that lasted only one hour total.

An additional challenge in this review was a frequent lack of detail provided for study and intervention protocols, rendering it difficult to assess the quality of individual study interventions. Future studies should include a sufficiently detailed protocol to allow readers to understand the scope of both experimental and control interventions, and to facilitate comparisons between intervention studies.

In terms of outcome measures, even within one type of outcome such as social functioning, some studies used standardized measures such as the SRS, whereas others used observation coding, for instance frequency of observed joy or initiation of engagement with the therapist. Over-reliance on parent-questionnaires is a frequent problem in intervention studies. Due to concerns about the accuracy and validity of parent-assessed ratings, it is advisable to at least have one assessment evaluated by a neutral third party in order to avoid rater bias (Glaser, Kronsnoble, & Warner Forkner, 1997).

In the present study sample, there is little information upon which to gauge demographic-related biases. The majority of studies were conducted in Western countries, leaving it unknown to what extent results can be generalized to different geographic or cultural settings. As only few studies report SES of participants' families, we cannot conclude on possible biases here. Studies in this sample generally did not examine potential effects of age or gender upon their results.

Future Directions

In consideration of the limitations identified above, future studies should aim for higher intensities and longer durations in order to generate more realistic and generalizable results. Additionally, studies should attempt to recruit larger samples and collect SES data from participating families. It is also important to match intervention groups on symptom severity and IQ in order to avoid potential confounds between intervention type and these parameters in the results. More rigorous baseline reporting, such as descriptions of symptom severity in each intervention group, is also needed. Another important addition is information about individuals who failed to complete the study, and consideration of potential biases or confounds resulting from attrition.

Additionally, while the present results generally support the idea that MI are as effective as NMI, and may have benefits above NMI, there is not enough information to dissociate which elements of music may be beneficial for specific outcomes. Some studies suggest that music can help ASD children control their moods and hence modulate their behavior (Devlin et al., 2008; Orr et al., 1998; Pasiali, 2004), but the present study did not find a significant difference between MI and NMI in terms of maladaptive behavior moderation. In fact, in one study, participants in the MI condition at least initially struggled with higher maladaptive behavior than NMI (Srinivasan et al., 2015). The lack of longitudinal investigations of maladaptive behaviors however does not allow for a definite conclusion. It is possible that other types of MI beyond the active music interventions reviewed here, such as AIT, could have such an effect (Dimitriadis & Smeijsters, 2011; Wan et al., 2010; Whipple, 2004). It is therefore reasonable to expect that different uses of music in the ASD population may lead to different outcomes. It is also possible that the strongest advantages of MI over NMI reside in factors other than those measured in this study. It has for instance been suggested that MIs create more joy and life satisfaction in ASD participants than alternative interventions (Bieleninik et al., 2017; Eren, 2015; Kim et al., 2009).

Conclusion

The present systematic review and meta-analyses generally support the effectiveness of MI, and provide evidence of benefits from MI as compared to NMI for social outcomes. However, methodological issues in the currently available literature impact the conclusiveness and generalizability of many study results. Of particular note, study sample sizes were small, and durations and intensities of interventions were considerably shorter than real-life scenarios. These results provide an up-to-date evaluation of the evidence for MI's benefits in children with ASD and identify recommended best practices to strengthen future clinical research in this domain.

Références

Accordino, R., Comer, R., & Heller, W. B. (2007). Searching for music's potential: A critical examination of research on music therapy with individuals with autism. *Research in Autism Spectrum Disorders, 1*(1), 101-115.

Allen, R., Hill, E., & Heaton, P. (2009). Hath charms to soothe...' An exploratory study of how high-functioning adults with ASD experience music. *Autism, 13*(1), 21-41.

American Music Therapy Association. (2013). Music Therapy. Retrieved from http://www.musictherapy.org/

American Psychiatric Association (Ed.) (2013). *DSM-5 - Manuel diagnostique et statistique des troubles mentaux (5)*. Philadelphia, PA.

Anderson, D. W., Kish, L., & Cornell, R. G. (1980). On stratification, grouping and matching. *Scandinavian Journal of Statistics*, 61-66.

Autism Speaks. (2018). What is Autism? Retrieved from https://www.autismspeaks.org/what-autism

Bean, J. L., & Eigsti, I.-M. (2012). Assessment of joint attention in school-age children and adolescents. *Research in Autism Spectrum Disorders, 6*(4), 1304-1310.

Bellini, S., & Hopf, A. (2007). The development of the Autism Social Skills Profile: A preliminary analysis of psychometric properties. *Focus on autism and other developmental disabilities, 22*(2), 80-87.

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: diagnostic validity. *The British Journal of Psychiatry, 175*(5), 444-451.

Bhat, A., & Srinivasan, S. (2013). A review of “music and movement” therapies for children with autism: embodied interventions for multisystem development. *Frontiers in Integrative Neuroscience, 7*(22). doi:10.3389/fnint.2013.00022

*[+]Bieleninik, Ł., Geretsegger, M., Mössler, K., Assmus, J., Thompson, G., Gattino, G., . . . Muratori, F. (2017). Effects of improvisational music therapy vs enhanced standard care on symptom severity among children with autism spectrum disorder: the TIME-A randomized clinical trial. *Jama, 318*(6), 525-535.

Bishop, D. (2013). *Children’s Communication Checklist (CCC-2)*: Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*: John Wiley & Sons.

Buday, E. M. (1995). The effects of signed and spoken words taught with music on sign and speech imitation by children with autism. *Journal of Music Therapy, 32*(3), 189-202.

Cogo-Moreira, H., de Avila, C. R. B., Ploubidis, G. B., & de Jesus Mari, J. (2013). Effectiveness of music education for the improvement of reading skills and academic achievement in young poor readers: a pragmatic cluster-randomized, controlled clinical trial. *PloS one, 8*(3), e59984.

Cohen, I., & Sudhalter, V. (1999). Pervasive Developmental Disorder Behavior Inventory (PDDBI-C). *NYS Institute for Basic Research in Developmental Disabilities*.

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale: SRS-2 Software Kit*: Western Psychological Services.

*+Crawford, M. J., Gold, C., Odell-Miller, H., Thana, L., Faber, S., Assmus, J., . . . Hassiotis, A. (2017). International multicentre randomised controlled trial of improvisational music therapy for children with autism spectrum disorder: TIME-A study. *Health technology assessment (Winchester, England), 21*(59), 1-40. doi:10.3310/hta21590

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., . . . Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics, 125*(1), e17-e23.

Devlin, S., Healy, O., Leader, G., & Reed, P. (2008). The analysis and treatment of problem behavior evoked by auditory stimulation. *Research in Autism Spectrum Disorders, 2*(4), 671-680.

Dimitriadis, T., & Smeijsters, H. (2011). Autistic spectrum disorder and music therapy: Theory underpinning practice. *Nordic Journal of Music Therapy, 20*(2), 108-122.

Dumont, E., Syurina, E. V., Feron, F. J. M., & van Hooren, S. (2017). Music Interventions and Child Development: A Critical Review and Further Directions. *Frontiers in Psychology, 8*(1694). doi:10.3389/fpsyg.2017.01694

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test*: American Guidance Service.

Eren, B. (2015). The use of music interventions to improve social skills in adolescents with autism spectrum disorders in integrated group music therapy sessions. *Procedia-Social and Behavioral Sciences, 197*, 207-213.

Fenson, L. (2007). *MacArthur-Bates communicative development inventories*: Paul H. Brookes Publishing Company Baltimore, MD.

Fletcher-Watson, S., & McConachie, H. (2017). The search for an early intervention outcome measurement tool in autism. *Focus on autism and other developmental disabilities, 32*(1), 71-80.

Fulton, E., Eapen, V., Črnčec, R., Walter, A., & Rogers, S. (2014). Reducing maladaptive behaviors in preschool-aged children with autism spectrum disorder using the Early Start Denver Model. *Frontiers in pediatrics, 2*, 40.

*+Gattino, G. S., Riesgo, R. d. S., Longo, D., Leite, J. C. L., & Faccini, L. S. (2011). Effects of relational music therapy on communication of children with autism: a randomized controlled study. *Nordic Journal of Music Therapy, 20*(2), 142-154.

Gebauer, L., Skewes, J., Westphael, G., Heaton, P., & Vuust, P. (2014). Intact brain processing of musical emotions in autism spectrum disorder, but more cognitive load and arousal in happy vs. sad music. *Frontiers in neuroscience, 8*, 192.

Geretsegger, M., Elefant, C., Mössler, K. A., & Gold, C. (2014). Music therapy for people with autism spectrum disorder. *Cochrane Database of Systematic Reviews*(6). doi:10.1002/14651858.CD004381.pub3

Geretsegger, M., Holck, U., Carpente, J. A., Elefant, C., Kim, J., & Gold, C. (2015). Common Characteristics of Improvisational Approaches in Music Therapy for Children with Autism Spectrum Disorder: Developing Treatment Guidelines. *Journal of Music Therapy, 52*(2), 258-281. doi:10.1093/jmt/thv005

*Ghasemtabar, S. N., Hosseini, M., Fayyaz, I., Arab, S., Naghashian, H., & Poudineh, Z. (2015). Music therapy: An effective approach in improving social skills of children with autism. *Advanced biomedical research, 4*, 157-157. doi:10.4103/2277-9175.161584

Glaser, B. A., Kronsnoble, K. M., & Warner Forkner, C. B. (1997). *Parents and Teachers as Raters of Children's Problem Behaviors* (Vol. 19).

Green, J., Charman, T., McConachie, H., Aldred, C., Slonims, V., Howlin, P., . . . Byford, S. (2010). Parent-mediated communication-focused treatment in children with autism (PACT): a randomised controlled trial. *The Lancet, 375*(9732), 2152-2160.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system: Manual*: American Guidance Service.

He, H., Yang, M., Duan, M., Chen, X., Lai, Y., Xia, Y., . . . Yao, D. (2018). Music Intervention Leads to Increased Insular Connectivity and Improved Clinical Symptoms in Schizophrenia. *Frontiers in neuroscience, 11*(744). doi:10.3389/fnins.2017.00744

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4): John Wiley & Sons.

Jackson, N. A. (2003). A survey of music therapy methods and their role in the treatment of early elementary school children with ADHD. In: American Music Therapy Association.

Jang, J., Dixon, D. R., Tarbox, J., & Granpeesheh, D. (2011). Symptom severity and challenging behavior in children with ASD. *Research in Autism Spectrum Disorders, 5*(3), 1028-1032.

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders, 34*(1), 81-86.

Kamioka, H., Tsutani, K., Yamada, M., Park, H., Okuizumi, H., Tsuruoka, K., Kitayuguchi, J. (2014). Effectiveness of music therapy: a summary of systematic reviews based on randomized controlled trials of music interventions. *Patient preference and adherence, 8*, 727.

*Katagiri, J. (2009). *The Effect of Background Music and Song Texts on the Emotional Understanding of Children with Autism* (Vol. 46).

*+Kim, J., Wigram, T., & Gold, C. (2008). The effects of improvisational music therapy on joint attention behaviors in autistic children: a randomized controlled study. *Journal of Autism and Developmental Disorders, 38*(9), 1758.

*Kim, J., Wigram, T., & Gold, C. (2009). Emotional, motivational and interpersonal responsiveness of children with autism in improvisational music therapy. *Autism, 13*(4), 389-409.

Kirschner, S., & Tomasello, M. (2010). Joint music making promotes prosocial behavior in 4-year-old children. *Evolution and Human Behavior, 31*(5), 354-364.

Kouklari, E.-C., Tsermentseli, S., & Auyeung, B. (2018). Executive function predicts theory of mind but not social verbal communication in school-aged children with autism spectrum disorder. *Research in developmental disabilities, 76*, 12-24.

Kover, S. T., Edmunds, S. R., & Weismer, S. E. (2016). Brief report: Ages of language milestones as predictors of developmental trajectories in young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 46*(7), 2501-2507.

LaGasse, A. B. (2014). Effects of a music therapy group intervention on enhancing social skills in children with autism. *Journal of Music Therapy, 51*(3), 250-275.

LaGasse, A. B. (2017). Social outcomes in children with autism spectrum disorder: a review of music therapy outcomes. *Patient related outcome measures, 8*, 23.

*+Lim, H. (2010). *Effect of "Developmental Speech and Language Training Through Music" on Speech Production in Children with Autism Spectrum Disorders* (Vol. 47).

*+Lim, H., & Draper, E. (2011). *The Effects of Music Therapy Incorporated with Applied Behavior Analysis Verbal Behavior Approach for Children with Autism Spectrum Disorders* (Vol. 48).

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*: Sage Publications, Inc.

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(6), 466-474.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2003). *Autism diagnostic observation schedule: ADOS*: Western Psychological Services Los Angeles, CA.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659-685.

Maloy, M., & Peterson, R. (2014). A meta-analysis of the effectiveness of music interventions for children and adolescents with attention-deficit/hyperactivity disorder. *Psychomusicology: Music, Mind, and Brain, 24*(4), 328-339. doi:10.1037/pmu0000083

Matson, J. L., Dempsey, T., LoVullo, S. V., & Wilkins, J. (2008). The effects of intellectual functioning on the range of core symptoms of autism spectrum disorders. *Research in developmental disabilities, 29*(4), 341-350.

Miendlarzewska, E. A., & Trost, W. J. (2014). How musical training affects cognitive development: rhythm, reward and other modulating variables. *Frontiers in neuroscience, 7*, 279.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine, 151*(4), 264-269.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational research methods, 11*(2), 364-386.

Orr, T. J., Myles, B. S., & Carlson, J. K. (1998). The impact of rhythmic entrainment on a person with autism. *Focus on autism and other developmental disabilities, 13*(3), 163-166.

Pasiali, V. (2004). The use of prescriptive therapeutic songs in a home-based environment to promote social skills acquisition by children with autism: Three case studies. *Music Therapy Perspectives, 22*(1), 11-20.

Pereira, A. M. (2007). Autismo infantil: Tradução e validação da CARS (Childhood Autism Rating Scale) para uso no Brasil.

Pickles, A., Le Couteur, A., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., . . . Byford, S. (2016). Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. *The Lancet, 388*(10059), 2501-2509.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raglio, A., Bellelli, G., Traficante, D., Gianotti, M., Ubezio, M. C., Villani, D., & Trabucchi, M. (2008). Efficacy of music therapy in the treatment of behavioral and psychiatric symptoms of dementia. *Alzheimer Disease & Associated Disorders, 22*(2), 158-162.

Rogers, S. J., Estes, A., Lord, C., Munson, J., Rocha, M., Winter, J., . . . Vismara, L. A. (2019). A multisite randomized controlled two-phase trial of the early Start Denver Model compared to treatment as usual. *Journal of the American Academy of Child & Adolescent Psychiatry*.

Samson, A. C., Hardan, A. Y., Lee, I. A., Phillips, J. M., & Gross, J. J. (2015). Maladaptive behavior in autism spectrum disorder: The role of emotion experience and emotion regulation. *Journal of Autism and Developmental Disorders, 45*(11), 3424-3432.

*Sandiford, G. A., Mainess, K., & Daher, N. (2012). *A Pilot Study on the Efficacy of Melodic Based Communication Therapy for Eliciting Speech in Nonverbal Children with Autism* (Vol. 43).

Särkämö, T., Tervaniemi, M., & Huotilainen, M. (2013). Music perception and cognition: development, neural basis, and rehabilitative use of music. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(4), 441-451.

Schopler, E., Reichler, R. J., DeVellis, R. F., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders, 10*(1), 91-103.

Schröder, C., Florence, E., Dubrovskaya, A., Lambs, B., Stritmatter, P., Vecchionacci, V., . . . Danion-Grilliat, A. (2015). Le modèle de Denver (early start Denver model). Une approche d'intervention précoce pour les troubles du spectre autistique. *Neuropsychiatrie de l'Enfance et de l'adolescence, 63*(5), 279-287.

*[+]Schwartzberg, E., & Silverman, M. (2013). *Effects of music-based social stories on comprehension and generalization of social skills in children with autism spectrum disorders: A randomized effectiveness study* (Vol. 40).

Seibert, J. M., Hogan, A. E., & Mundy, P. C. (1982). Assessing interactional competencies: The early social-communication scales. *Infant Mental Health Journal, 3*(4), 244-258.

*+Sharda, M., Tuerk, C., Chowdhury, R., Jamey, K., Foster, N., Custo-Blanch, M., . . . Hyde, K. (2018). Music improves social communication and auditory–motor connectivity in children with autism. *Translational psychiatry, 8*(1), 231.

*+Simpson, K., Keen, D., & Lamb, J. (2015). Teaching receptive labelling to children with autism spectrum disorder: A comparative study using infant-directed song and infant-directed speech. *Journal of Intellectual & Developmental Disability, 40*(2), 126-136. doi:10.3109/13668250.2015.1014026

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1998). *Vineland social-emotional early childhood scales: Manual*: Pearson Assessments.

Sparrow, S. S., & Cicchetti, D. V. (1985). Diagnostic uses of the vineland adaptive behavior scales. *Journal of Pediatric Psychology, 10*(2), 215-225.

*+Srinivasan, S. M., Eigsti, I.-M., Gifford, T., & Bhat, A. N. (2016b). The effects of embodied rhythm and robotic interventions on the spontaneous and responsive verbal communication skills of children with Autism Spectrum Disorder (ASD): A further outcome of a pilot randomized controlled trial. *Research in Autism Spectrum Disorders, 27*, 73-87.

*Srinivasan, S. M., Eigsti, I.-M., Neelly, L., & Bhat, A. N. (2016a). The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with Autism Spectrum Disorder (ASD): A pilot randomized controlled trial. *Research in Autism Spectrum Disorders, 27*, 54-72.

*Srinivasan, S. M., Park, I. K., Neelly, L. B., & Bhat, A. N. (2015). A comparison of the effects of rhythm and robotic interventions on repetitive behaviors and affective states of children with Autism Spectrum Disorder (ASD). *Research in Autism Spectrum Disorders, 18*, 51-63.

Suthers, L., & Niland, A. (2007). An exploration of young children's engagement with music experiences. *Listen to their voices: Research and practice in early childhood music*, 19-32.

Tang, H.-Y. J., & Vezeau, T. (2010). The use of music intervention in healthcare research: a narrative review of the literature. *Journal of Nursing Research, 18*(3), 174-190.

*+Thompson, G., McFerran, K., & Gold, C. (2013). *Family-centred music therapy to promote social engagement in young children with severe autism spectrum disorder: a randomized controlled study* (Vol. 40).

Tsermentseli, S., Tabares, J. F., & Kouklari, E. C. (2018). The role of every-day executive function in social impairment and adaptive skills in Autism Spectrum Disorder with intellectual disability. *Research in Autism Spectrum Disorders, 53*, 1-6.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods, 7*(1), 55-79.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

Vismara, L. A., & Rogers, S. J. (2010). Behavioral treatments in autism spectrum disorder: what do we know? *Annual review of clinical psychology, 6*, 447-468.

Wan, C. Y., Demaine, K., Zipse, L., Norton, A., & Schlaug, G. (2010). From music making to speaking: engaging the mirror neuron system in autism. *Brain research bulletin, 82*(3-4), 161-168.

Warren, Z., McPheeters, M. L., Sathe, N., Foss-Feig, J. H., Glasser, A., & Veenstra-VanderWeele, J. (2011). A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics, 127*(5), e1303-e1311.

Whipple, J. (2004). Music in intervention for children and adolescents with autism: A meta-analysis. *Journal of Music Therapy, 41*(2), 90-106.

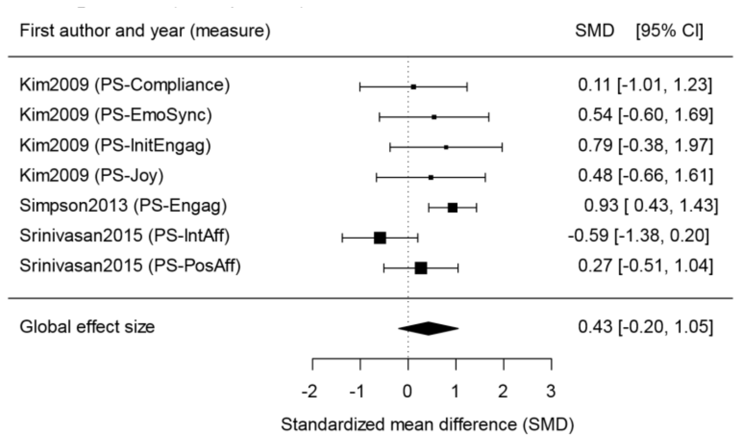
White, S. W., Keonig, K., & Scahill, L. (2007). Social skills development in children with autism spectrum disorders: A review of the intervention research. *Journal of Autism and Developmental Disorders, 37*(10), 1858-1868.

* studies included in systematic review

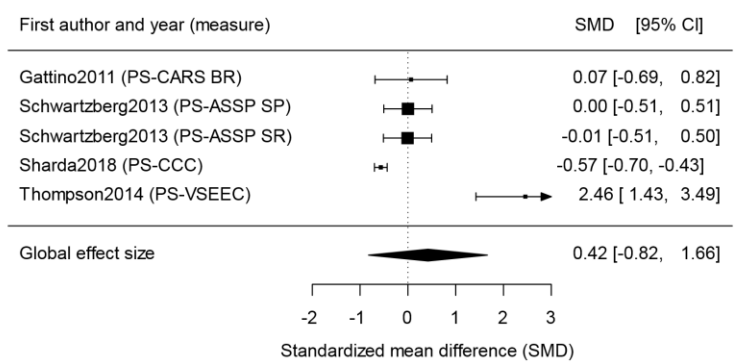
+ studies included in meta-analyses

Appendix A

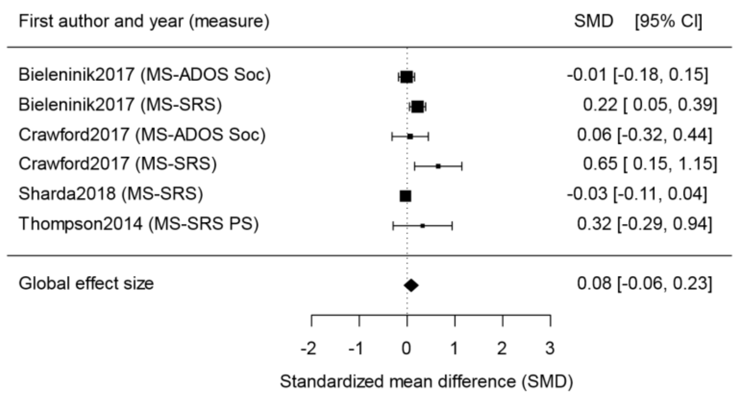
Forest plot for the meta-analysis comparing MI vs NMI for prosocial measures in within-session studies



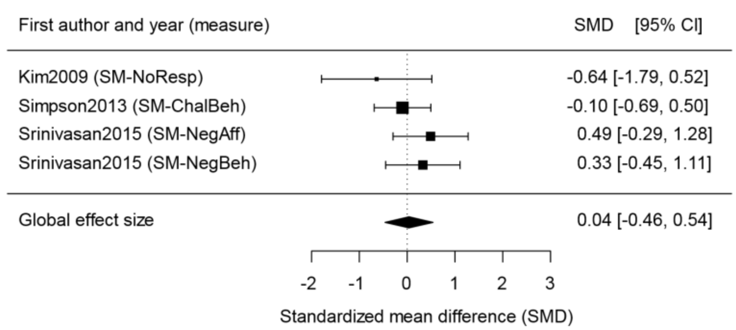
Forest plot for the meta-analysis comparing MI vs NMI for prosocial measures in longitudinal studies



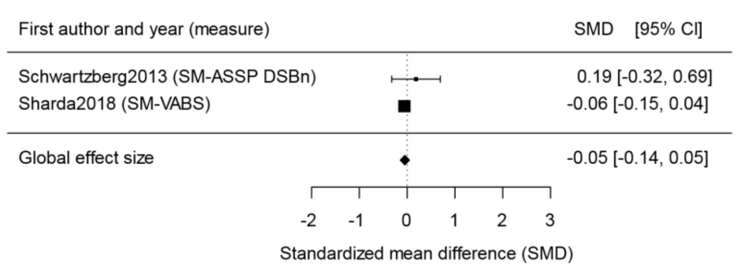
Forest plot for the meta-analyses comparing MI vs NMI for mixed social measures in longitudinal studies



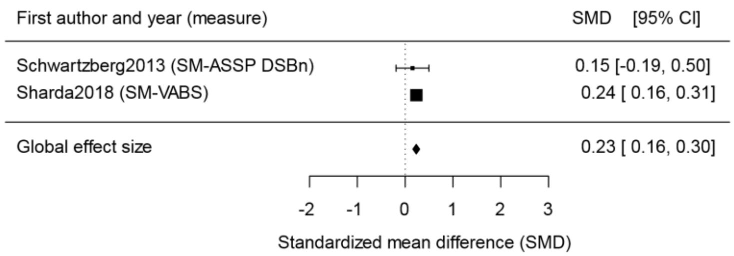
Forest plot for meta-analysis comparing MI vs NMI for socially maladaptive behavior measures in within-session studies



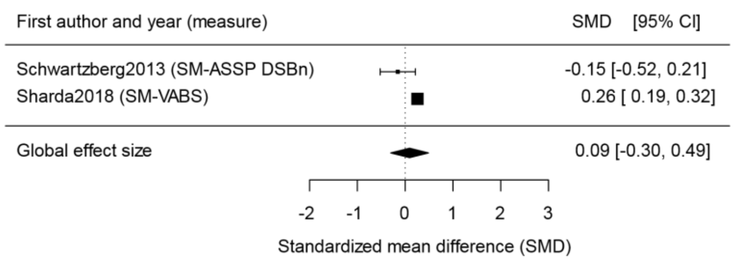
Forest plot for meta-analysis comparing MI vs NMI for socially maladaptive behavior measures in longitudinal studies



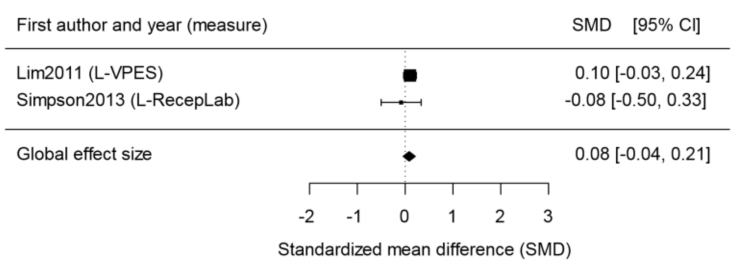
Forest plot for meta-analysis of the effect of MIs on socially maladaptive behavior measures in longitudinal studies



Forest plot for meta-analysis for the effect of NMIs on socially maladaptive behavior measures in longitudinal studies



Forest plot for meta-analysis comparing MI vs NMI for language measures in within-session studies



Forest plot for meta-analysis comparing MI vs NMI for language measures in longitudinal studies