System 124 (2024) 103350

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

Review

Investigating the substantive linguistic effects of using songs for teaching second or foreign languages to preschool, primary and secondary school learners: A systematic review of intervention research

Catherine Hamilton *, Johannes Schulz, Hamish Chalmers, Victoria A. Murphy

University of Oxford, Department of Education, 15 Norham Gardens, Oxford, OX2 6PY, UK

ARTICLE INFO

Keywords:
Systematic review
Songs
Music
Foreign language learning
Second language learning
Young language learners
Intervention research

ABSTRACT

Songs are popular resources with teachers of young language learners. In addition to important socioemotional and developmental outcomes, a common assumption is that songs will help support learning the target language. This systematic review narratively synthesises evidence from intervention research on the effects of using songs in second or foreign language classrooms on linguistic outcomes among children aged 2–18 years. 1862 potentially relevant reports were identified. After screening, 60 intervention studies from 23 countries were located that assessed the relationship between using songs in the classroom and substantive linguistic outcomes. These were vocabulary acquisition, grammatical learning, and speaking, listening, reading, and writing skills. While most of the assembled literature made positive causal claims about the relationship between singing songs and these outcomes, a majority were not appropriately designed to support these claims. Our formal assessment of the robustness of the designs and other methodological characteristics of the included studies suggests that it is not possible to draw firm causal inferences about the effect of using songs on linguistic outcomes. This systematic review makes the case for conducting further robustly designed intervention research to better inform our understanding of the linguistic effects of using songs to teach young language learners.

1. Introduction

This systematic review synthesises evidence from intervention research investigating the use of songs as pedagogical tools with young second or foreign language learners (YLLs) aged between two and 18 years in formal educational contexts (i.e., preschool, primary and secondary schools). Songs are popular resources with YLL teachers worldwide (Linse, 2006; Şevik, 2011). Distinct from chants or rhymes, which have salient rhythm but not melody (Davis & Fan, 2016; Forster, 2006), songs are sometimes conflated into 'musical activities' – a category that also includes reading song books, creating instruments, and listening or dancing to instrumental music or songs (Paquette & Rieg, 2008). Teachers present songs as individual, small-group or whole-class singing and listening activities via screen, audio recording, or live performance (Hamilton & Murphy, 2023), and as written lyrics for gap-filling, sequencing,

* Corresponding author.
E-mail addresses: catherine.hamilton@education.ox.ac.uk (C. Hamilton), johannes.schulz@education.ox.ac.uk (J. Schulz), hamish.chalmers@education.ox.ac.uk (H. Chalmers), victoria.murphy@education.ox.ac.uk (V.A. Murphy).

https://doi.org/10.1016/j.system.2024.103350
Received 30 April 2023; Received in revised form 18 May 2024; Accepted 21 May 2024
Available online 31 May 2024

grammar or vocabulary exercises, or stimuli for creative output (Davanellos, 1999; Walker, 2006). In this paper, 'songs' encapsulates all pedagogical uses teachers make of songs containing lyrics (i.e., not purely instrumental music), while 'music' encompasses both songs and instrumental activities.

In a survey of 4696 English language teachers from 144 countries, 67% of respondents reported that they used songs often or every lesson (Garton et al, 2011). A survey conducted among 270 schools in Ireland (Harris and O'Leary, 2009) asked teachers to rank 18 foreign language teaching/learning activities in order of pupil enjoyment and frequency of use. 'Raps/songs' were ranked second in terms of enjoyment and eighth in terms of frequency of use (p.5). Teachers often believe using songs for language teaching has educational and linguistic benefits, particularly with younger learners (Hamilton & Murphy, 2023). Notwithstanding the inevitably contextually specific samples in these studies, they nonetheless provide evidence that songs are well regarded among language teachers as pedagogical activities, although they are not perhaps used as frequently as their popularity suggests they might be.

There are many different reasons why songs might be used in FL teaching, and many associated outcomes that teachers and learners value. For example, singing songs has been associated with supporting social and emotional development (for a review see Váradi, 2022) and songs observed for their capacity to engage and motivate learners (e.g., Kaminski, 2016). In addition, songs are often believed to support the learning of the target language itself. While we recognise the potential value of singing songs with young learners for a range of purposes, it is their effects on this latter outcome that forms the specific focus of this review.

1.1. Epistemological orientation of the review

The use of songs in FL teaching can and has been investigated using a variety of designs, following a variety of epistemological traditions. For example, Kaminski (2016) and Geisler (2008) gathered longitudinal qualitative data to investigate the use of songs in FL teaching in German primary schools. Both studies found that motivation and engagement for learning English improved over the period of observation. They provide rich, contextualised information and contribute important and meaningful evidence to our understanding of the use of songs in FL teaching and learning. However, such studies are not well suited for understanding the casual relationships between singing songs and *linguistic* outcomes, as we explain in the following.

This review seeks to address the question of whether we have reliable evidence upon which we can draw firm conclusions about the causal relationships between using songs in FL lessons and their effects on substantive FL learning outcomes, in particular vocabulary acquisition, grammatical learning, and speaking, listening, reading and writing skills. We are concerned with the question of causality, since there is a clear instinctual belief among teachers that using songs is beneficial for improving FL learning outcomes, and specifically linguistic outcomes. We situate our investigation into causality within the epistemological tradition of experimentation in social sciences research laid out by Campbell (1957), who reminded the field that "the very minimum of useful scientific information involves at least one formal comparison and therefore at least two careful observations" (Campbell, 1957:298). For the purposes of this review, then, we are concerned only with intervention research. That is, studies where a teaching approach involving songs or rhythm-salient input is implemented, and the linguistic outcomes of students measured to establish the effects of using songs on those outcomes. Primarily, this means experimental and quasi-experimental designs.

Designing studies to establish causality through such comparative methods in a practical educational setting is not always straightforward and can be challenging. Consequently, there are a range of interventional research approaches taken which address these practical challenges whilst making valuable contributions to educational research. Campbell and colleagues (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al, 2002) and those that continue the tradition in the social and educational sciences (e.g., Connolly et al., 2017; Gorard, 2003, 2013; Slavin, 1986) identify designs that are more or less robust in terms of their capacity to confidently identify causal relationships, should they exist. At the first point of the methodological scale is an approach that involves a single group of participants: they engage with the approach under investigation and their performance before and after is compared. This design gives an indication of potential effects, but with no formal comparison it is impossible to estimate what would have happened had they not been taught with that approach. A more robust approach is to add a control group, and this can be done in several ways. The simplest is to compare one class against another, but this creates challenges for detecting causality because it is impossible to establish with certainty whether any differences in outcomes between groups is a result of the intervention or because of existing differences in the average characteristics of participants in each class (e.g., different levels of prior attainment). The comparison process can be made more robust by using statistical matching of participants in an attempt to ensure that comparison groups are fair approximations of each other. However, statistical matching can only account for characteristics that researchers know about and can measure. Therefore, the use of random allocation to comparison groups is considered to be particularly robust in ensuring that allocation bias (Nunan, Henghan, & Spencer, 2018) is minimised, that groups are unbiased approximations of each other, and that differences in outcomes between groups at the end of a study can therefore be more confidently attributed to the intervention rather than to systematic differences in the characteristics of the groups being compared.

We recognise that all of these designs have been used at one time or another in investigating casual relationships between using songs and FL learning outcomes, thus we adopt a 'best evidence synthesis' approach (Slavin, 1986) for this review. That is, we aim to ascertain what can be concluded from intervention research in the field, and will consider the relative robustness of the body of evidence in doing so. Principally, this will be addressed by assessing the methodological quality of the body of literature against a tool designed for this purpose (see section 3.8). As Slavin (1986:10) states, "a best-evidence synthesis should produce and defend conclusions based on the best available evidence, or in some cases may conclude that the evidence currently available does not allow for any conclusions." In the absence of designs that can reliably draw causal links between using songs and language learning outcomes, we will describe what can be concluded from the current state of our knowledge on the topic, and what areas need to be built on with more robust designs.

Our hope is that by assembling, describing, and evaluating intervention research, we will provide important information to be used alongside other, equally important information about the use of songs derived from other research traditions, so that teachers can make informed decisions about their practice and its likely effects on their students' linguistic outcomes.

2. Background and rationale

Teachers commonly express strong intuitions that songs 'work' for a variety of educational purposes, including memorisation of concepts or vocabulary, improving pronunciation, establishing grammatical knowledge, supporting classroom routines and behaviour, and motivating learners (Davanellos, 1999; Forster, 2006; Hamilton & Murphy, 2023; Paquette & Rieg, 2008; Saricoban & Metin, 2000; Schoepp, 2001; Walker, 2006). There is clear anecdotal support from practitioners for using songs to achieve linguistic amongst other FL outcomes.

However, given their popularity, there is a surprising lack of robust empirical evidence supporting the use of songs to achieve linguistic development with YLLs (Davis, 2017; Degrave, 2019; Engh, 2013; Sposet, 2008; Werner, 2020). Davis' (2017) 'critical review' only identified nine classroom intervention studies from eight countries seeking evidence for using songs with 3–to–12-year-olds, and a further six that were removed upon screening due to insufficient reporting of their interventions or measures. Three included studies involved an external researcher conducting a workshop or lesson incorporating songs, and five studies involved the class teacher using songs in regular lessons. Outcomes included receptive and productive vocabulary, motivation, and pronunciation, with six studies focusing on vocabulary acquisition. There were equivocal findings for the effect of songs on vocabulary acquisition. Since songs (or rhythm-salient input in one case) were only isolated as a variable in three of the included studies, any effects on linguistic outcomes cannot reliably be attributed to songs alone. With a small sample of studies with heterogenous participant demographics, methodologies, and outcome measures, Davis concluded that overall substantive effects of using songs for language outcomes were tentatively positive, but still ambiguous. However, Davis (2017) searched for combinations of 'young learners', 'songs' and 'music' and may have missed relevant studies with other keywords, thus pointing us towards taking a more systematic and replicable approach in future reviews.

Finding similarly sparce material for the period 1937–2007, Sposet (2008) conducted a 'bibliographical review' of research, reporting that 15 of 23 included studies found positive outcomes for using music for second language acquisition (SLA) with learners from kindergarten through to adulthood. Sposet states that the scant available evidence does not support firm conclusions about music's role in SLA. Sposet also claims that the included data appear to show music's positive effect on SLA, particularly pronunciation, but this does not appear to be fully supported by the review's findings. Werner (2020) conducted a more recent 'research synthesis' investigating classroom-based intervention studies where lyrics-based language instruction was assessed for potential advantages or costs to linguistic outcomes among learners aged from primary (earliest reported age is 7 years) to university levels. Studies without control groups were excluded, thus 28 classroom intervention studies were included in the final analysis. Werner found a positive overall effect of lyrics-based instruction for English vocabulary acquisition and verbal recall, but scant research investigating target languages other than English or other linguistic outcomes. The prior reviews of evidence in this area do not report replicable, transparent and systematic methods, and formal study quality appraisal is absent. The reviews by Davis (2017), Sposet (2008), and Werner (2020) thus leave us unable to draw firm or meaningful conclusions about the substantive linguistic effects of using songs to teach YLLs since bias cannot be evaluated without quality appraisal of included studies. Overall, then, a transparent, systematic, and replicable approach to evaluating the state of the knowledge is needed.

In the remainder of this section we briefly introduce empirical evidence on using songs with young learners in L1 classrooms, then highlight three often-cited theoretical motivations for research investigating songs with YLLs in L2 contexts, and finally introduce evidence about songs' involvement in speech and language development gathered from transdisciplinary studies with infants. The section concludes with the research questions for this systematic review.

2.1. Empirical evidence for using songs in L1 classrooms

A handful of classroom studies provide equivocal support for songs' role in preschoolers learning L1 English vocabulary (Crosswhite, 1996; Joyce, 2011) or phonological awareness (Lehman, 2019) but they are statistically underpowered, and in any case not generally applicable to L2 YLL contexts because the participants are learning their L1 in naturalistic contexts, not an L2 in input-limited contexts. A review of early years music-making studies (Lonie, 2010) found ambiguous support for claims made about music's transfer effects to L1 literacy development, echoing the meta-analyses discussed in 2.2 below.

2.2. Theoretical motivations in existing L2 research

There is a lack of well-grounded theoretical motivation for research conducted into using songs with YLLs. Teachers tend not to question why songs 'work' for multiple educational purposes, including children's language, behaviour, social and concept-knowledge development (Hamilton & Murphy, 2023). This circulation of 'folk theory' (Bruner, 1996) amongst practitioners in turn influences research being conducted and published in peer-reviewed journals. Bruner (1996) notes that teachers' subjective beliefs and tacit knowledge are often afforded a similar status to scientifically tested hypotheses because they stand the test of public scrutiny over time, solidifying past conjecture into received wisdom. Hamilton and Murphy (2023) found such 'folk theory' about songs' influence on YLLs' linguistic outcomes often goes unchallenged in journal publications, reinforcing cultural beliefs that music and songs confer 'transfer benefits' between cognitive or academic domains. When more carefully considered research is conducted, the picture is less

clear. Recent meta-analyses exploring evidence on the putative benefits of music training on cognitive and academic outcomes found mixed results. Controlling for study quality removed demonstrable or consistent effects of general music training on children's mathematic or literacy skills (Sala & Gobet, 2020) or produced a small amount of reliable evidence that learning to play an instrument during the school years has a modest but significant impact on cognitive or academic outcomes (Román-Caballero, Vadillo, Trainor & Lupiáñez, 2022). These findings challenge long-held beliefs in the 'Mozart effect' (Rauscher, Shaw, & Ky, 1993), a 'scientific legend' that has captured news headlines and teachers' attention for decades (Bangerter & Heath, 2004). Despite the scarcity of credible evidence, beliefs that music training makes you more intelligent and confers academic benefits extrinsic to music persist, particularly in relation to language learning (see, for example, the volume exploring rhythm, melody and cognition in language education edited by Fonseca-Mora et al, 2015).

Engh (2013) reviewed theoretical support for using songs to teach languages, citing a selection of transdisciplinary material including anthropological, cognitive, and pedagogical research. However, Hamilton and Murphy (2023) found limited substantiating evidence for these diverse theoretical claims, identifying that experiential and experimental evidence are circulated uncritically, and with increasing enthusiasm, in a liminal space between teacher-facing non-peer-reviewed publications (e.g., blogs, publications from ELT special interest groups, and textbooks) and peer-reviewed research literature. The following three subsections review key theoretical foundations proposed in the literature to justify using songs to achieve linguistic outcomes with YLLs.

2.2.1. *Involuntary mental rehearsal*

Krashen (1983) hypothesised that after one or two hours of input, FL words echo spontaneously in learners' heads in a form of spontaneous playback. He suggests that this involuntary mental rehearsal permits learners to speak their target language more confidently and fluently, even after a decade of not using the language. Krashen based his hypothesis on an anecdote about German "rattling in [his] brain" (Krashen, 1983:42) during a German conference, and Barber's (1980) account of having a "rising din of Russian in [her] head" (cited in Krashen, 1983:42), prompting the name 'din' hypothesis. In follow-up questionnaire studies with high school and university languages students (Bedford, 1985; Guerrero, 1987; Parr & Krashen, 1986), participants confirm they identify with the din only *after* reading a description of the phenomenon, arguably leading them to answer affirmatively. There is little empirical data regarding the din's supposed "real practical value" (Krashen, 1983:44), psycholinguistic workings, or applicability across learner demographics.

A frequently cited source of evidence for songs' language education benefits which takes Krashen's (1983) 'din' hypothesis as its theoretical cornerstone is Murphey's (1990) paper extolling the mnemonic benefits of songs. Murphey (1990:53) administered his students ($n = 49$) a "tentative pilot questionnaire" to see if they, like him, experienced involuntary mental song rehearsal. All respondents identified with his experience. Based on this survey, Murphey promulgated what he called the 'song stuck in my head' (SSIMH) phenomenon. Murphey (1990) did not claim that SSIMH has proven educational benefits, just that it *may* prove to be advantageous for language learning by activating the 'LAD' (Language Acquisition Device; Chomsky, 1965). Murphey called for further research into what he considered an interesting idea, echoing similar calls from Bedford (1985) and Guerrero (1987). Despite Murphey's reticence about the SSIMH's evidential foundations, the lack of empirical evidence supporting 'din' (Krashen, 1983), and the "opaque black box" (Mitchell, Myles, & Marsden, 2019:55) that is the inner workings of the LAD, songs' mnemonic and consequent linguistic benefits for SLA are often stated (Davanellos, 1999; Degrave, 2019; Fonseca Mora, 2000; Thain, 2010) based on these unfalsified hypotheses.

A recent theoretical exploration of 'earworms' (Arthur, 2023) indicates that there is nascent evidence of songs that are easier to sing along to (e.g., *Baby Shark*) being rehearsed involuntarily, and that the phonological loop is activated during subvocal articulation, which may point towards linguistic benefits of harnessing such earworms. However, there is no consensus on what features of songs make them "stick" or how to achieve an earworm deliberately, since a key characteristic is their involuntary intrusion into the mind. There is thus still some way to go before the phenomenon of involuntary mental rehearsal is thoroughly understood and reliably linked to substantive linguistic outcomes in classroom contexts, or the precise 'dose' of songs to achieve optimal involuntary mental rehearsal is discovered. It is clear, however, that the phenomenon has captured teachers' interest and that researchers can draw upon more recent (and potentially more robust) evidence than Krashen (1983) and Murphey (1990) when discussing such theory as motivation for their investigations into the effects of songs on linguistic outcomes.

2.2.2. *Musical intelligence and learning styles*

Practitioners and researchers invoke musical intelligence and learning styles research (often without noting that these are discrete research fields) as theoretical foundations for using songs, typically with limited supporting evidence or critique. Fonseca-Mora (2000) cites Gardner's (1983) theory of multiple intelligences, advocating a variety of activities for different learners and emphasising musical intelligence's relevance in language teaching. However, it lacks empirical evidence that would support citing learning styles and musical intelligence specifically as reasons for using songs in language lessons. Engh's (2013) widely cited theoretical review builds on Fonseca-Mora (2000), without fully critiquing the earlier paper, linking learner styles, multiple intelligences, and motivation, claiming that addressing learners' preferred auditory styles or musical intelligence directly by learning English through music increases their motivation. This demonstrates how, without the addition of further studies focused on determining causality, the weight of published research can build increasing certainty without a firm base.

Critical appraisal of learning styles and multiple intelligences research is essential, since both fields have been criticised, with meta-analyses finding them incoherent, self-interested, and without replicable, rigorous findings (Coffield, Moseley, Hall & Ecclestone, 2004; Waterhouse, 2006). There is a lack of causal evidence to support linking learner preferences to pedagogy (Coffield et al., 2004). Claims that using songs supports learners' preferred learning styles, and hence promotes language learning, are therefore

unsubstantiated. Recent correlational studies in neuroscience have potentially reopened the case for the existence of multiple intelligences (Shearer, 2020), but it remains to be seen whether increased attention to musical intelligence through classroom musical activities facilitates language learning.

2.2.3. Prosodic bootstrapping hypothesis

Another theoretical foundation for empirical research could be the prosodic bootstrapping hypothesis (PBH; Gleitman, 1990; Morgan & Demuth, 1996). In L1 acquisition, prosody represents children's first encounter with linguistic input *in utero* (see evidence review in Gervain, Christophe, & Mazuka, 2020). According to PBH, prosody scaffolds language learning, helping children parse input through the prosody-based prominence of content words (nouns/verbs) and mapping these onto salient visual objects, assisting acquisition of L1 lexical and morphosyntactic features. There is evidence that prosodic bootstrapping assists YLLs in classrooms with L2 word-order acquisition and elicited imitation tasks (Campfield & Murphy, 2013, 2014), and adults in lab-based studies (Saksida, Flo, Guedes, Nespor & Garay, 2021). Teaching L2 prosody and suprasegmental features explicitly may improve fluency and comprehensibility of L2 learners' speech (Gordon & Darcy, 2016). It could be the case that presenting YLLs with L2 input through singing or chanting, where prosody is especially salient, enhances the L2 learning process relative to other modes of oral presentation, such as hearing conversational speech or reading prose aloud.

2.2.4. Summary of reviewed theoretical approaches

In summary, involuntary mental rehearsal as it is presented by the 'din' hypothesis or Song Stuck in My Head phenomenon, and musical intelligence or learning styles, are popular theoretical refrains in justifying using songs as YLL pedagogy. However, as discussed above, they provide unstable foundations for experimental work to build upon. The Prosodic Bootstrapping Hypothesis provides stronger theoretical motivation for research investigating songs' influence in FL learning since it brings an empirically tested L1 theory into the instructed YLL domain.

2.3. Transdisciplinary evidence

In addition to Campfield and Murphy's (2013; 2014) theoretical contribution that PBH may assist L2 acquisition and can be facilitated by prosodically salient input such as rhymes, tangential findings from transdisciplinary studies suggest pursuing evidence for using songs with YLLs is worthwhile. Investigations into infant cry melodies propose that musical elements of infant pre-speech are a necessary stage in language acquisition, not a by-product (Wermke & Mende, 2009, 2016): their Melody-Development Model is a complexity hypothesis where infants' early vocal melodies iteratively develop into more complex combinations and phases of pre-speech and speech, relating infant cries to singing. There is cumulative evidence about infant-directed singing's affective importance and lullabies' universality (Bainbridge, Youngers, Bertolo, Atwood, Lopez, Xing & Mehr, 2021; Trehub & Trainor, 1998; Trehub et al., 1993). Combined with recent insights into how early exposure to singing (as opposed to general or background music) contributes to early speech and language development (Franco, Suttora, Spinelli, Kozar, & Fasolo, 2021), infant studies indicate that evidence for using songs in YLL education is worth seeking.

2.4. Summary and research questions

Although teachers have scant robust evidence underpinning intuitions that songs are an effective language-learning tool, their experiential wisdom merits careful analysis of research literature that may support pedagogical choices. As Paran (2017) argues, intuition and research are not competing foundations for teaching practice. Indeed, conceptions of evidence-based practice explicitly acknowledge the importance of considering practitioner experience and expertise alongside the best available external evidence when making choices about practice (Chalmers, 2016). Building on intuition, experience, expertise, *and* external research findings avoids teaching becoming "merely the transmission of self-perpetuating, unsupported beliefs and prejudices" (Paran, 2017:506). Currently, no demonstrable consensus exists within the literature on what the substantive linguistic effects of using songs in children's L2 education might be. Without access to empirical evidence from reliable sources, teachers risk basing practice on unexamined intuition and overlooking approaches that would best support YLLs. We hope that our review, in addressing the research questions below, contributes useful substantive evidence of what is known about songs' effectiveness as pedagogical tools for teaching YLLs and serves as a solid foundation for future primary research to build on.

RQ1: What is the extent and nature of intervention research investigating the substantive linguistic effects of using songs to teach second or foreign languages to young learners in formal education contexts?
RQ2: What can be reliably concluded from intervention research identified in RQ1 about the effects of using songs with young language learners on substantive foreign language learning outcomes?

3. Methodology

3.1. Protocol and registration and reporting standard

This review is reported in line with the standards laid out by the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Page et al., 2021). While initially formulated to guide the reporting of systematic reviews in

healthcare, PRISMA is widely regarded as appropriate (and widely adopted) for use in any discipline. This includes the social sciences generally (Gough et al., 2012), and education (Zawacki-Richter et al., 2020) and applied linguistics (Csizér et al., 2022) specifically.

The review protocol was written using the PRISMA extension for protocols (PRISMA-P; Moher et al., 2015) and registered prospectively on the International Database of Education Systematic Reviews (IDESR) in December 2021, under registration number IDESR000017 (https://idesr.org/article/IDESR000017).

3.2. Eligibility criteria

Table 1 presents the eligibility criteria. Published papers and grey literature in any language were included to seek all available evidence dealing with typically developing language learners in preschool, primary and secondary school contexts worldwide.

3.3. Information sources

Table 2 shows the consulted databases in education, linguistics, psychology, and multidisciplinary research. The authors speak French, German and Spanish, so relevant databases in these languages were included to broaden the search. After consulting with research librarians at the University of Oxford and linguistics colleagues in Europe, these databases were chosen because they provide meta-catalogues of university libraries, human and social sciences databases, and grey literature produced in French, German and Spanish. All databases accept Boolean search syntax. There were differing limits to how many search terms could be included, as reflected in the search strings reported in Appendix A.

Table 1
Eligibility criteria.

Item	Inclusion criterion	Rationale
Bibliographic information	Include 1: Studies with a full reference or sufficient information. Exclude 1: Studies with insufficient bibliographic information.	Without sufficient bibliographic information, retrieval of works is unfeasible.
Date of publication	Include 2: Published on any date.	Attempting to collect all eligible studies regardless of date of publication.
Participants	Include 3: Studies on typically developing foreign language learners. Include studies even if no explicit reference is made to learning ability if reasonable assumption can be made that participants are comprised mainly of typically developing individuals. Exclude 3: Studies that exclusively target non-typically developing learners or learners with Developmental Language Disorder.	This review seeks to assess effects of songs as a pedagogical tool in typically developing school populations. The findings for non-typically developing populations may not generalise to a larger population, thus such results will not be extrapolated or included in this review.
	Include 4: Studies conducted in preschool, primary or secondary schools (students aged 2–18) or other formal settings (e.g., playgroups, after-school clubs) worldwide. Exclude 4: Studies conducted in university, or adult educational contexts; informal settings (e.g., at home).	This study focuses on the outcomes of using songs for learners in formal contexts between age 2 and 18, since adult learners (over 18) have different learning capacities and educational goals. Findings from studies conducted in informal settings may not generalise to formal educational settings, thus such results will not be extrapolated or included in this review.
Intervention	Include 5: Studies where singing songs, choral chanting, or nursery rhymes are included as a whole-class or group activity. Exclude 5: Studies where musical instruments are the intervention focus, not singing, chanting or nursery rhymes.	This review focuses on the linguistic outcomes of using songs as pedagogical tools, thus the intervention must include songs with words, not purely an instrumental intervention (e.g., whole-class ukulele lessons).
Outcomes	Include 6: Primary research studies reporting any measure of language acquisition including but not limited to vocabulary, grammar or phonology outcome measures. Include studies that report either quantitative or qualitative measures of outcomes. Exclude 6: Systematic reviews or studies that provide only narrative evaluation of an intervention but do not include outcome measures of language acquisition including vocabulary, grammar or phonology; studies that measure only non-language outcomes, e.g., satisfaction, happiness, engagement.	A synthesis of empirical findings in this field of literature is impossible without the reporting and evaluation of concrete data.
	Include 7: Any type of study design that attempts to identify a causal relationship. Exclude 7: Studies where no attempt to identify causality is made (e.g., ethnographies, observations)	Given the expected scarcity of research in this area, we take an inclusive approach to study types designed to identify causality.
Publication status	Include 8: Grey literature. Exclude 8: Do not exclude studies based on publication status.	This paper seeks to offset potential publication bias by including a wider range of research, including grey literature.
Language of publication	Include 9: Studies published in any language. Exclude 9: Do not exclude studies based on the language of publication.	Limiting this review to studies published in English may result in a systematic neglect of a particular body of research.

Table 2
List of databases.

Discipline	Database			
	English	German	French	Spanish
Education	ProQuest Education Collection (including ERIC), British Education Index EBSCO; Education Abstracts (H.W. Wilson)	Fachportal Pädagogik	n/a	n/a
Linguistics	ProQuest Linguistics Collection (including LLBA); MLA International Bibliography	n/a	n/a	n/a
Psychology	PsychInfo	PsynDEX	n/a	n/a
Multidisciplinary	Web of Science, Scopus	Humboldt University Berlin; Center for Research Libraries Global Resources Network (CRL)	Cairn.info; SUDOC; Pascal-Francis; CRL	CRL
Grey literature	ProQuest Dissertations & Theses Global; OpenGREY; EthOS	n/a	theses.fr	TESEO educacion.gob.es

3.4. Search strategy

The main search strategy for ProQuest (see Table 3) was developed iteratively to balance sensitivity with specificity. Pilot searches, which included participants' ages (e.g., "5 year* old*" or "five year* old*" or "aged 5″ or "aged five"), returned excessive irrelevant results about child language disorders. Therefore, settings were specified rather than participants' ages. The original intervention part of the string (intervention OR RCT OR "randomi?ed control*" OR research OR "action research" OR study) returned many irrelevant results about medical interventions when piloted. We limited searches to the musical nature of the intervention, not type of study design, and instead applied design during the selection process.

All search terms were included in the ABSTRACT search frame on English searches, as piloting indicated this returned the most relevant results. Finally, search terms were translated and cross-referenced to check their accuracy in relevant French, German and Spanish journals. Placement of search terms in the title, abstract or full text varied across languages and databases. Piloting was conducted to ensure we captured maximum relevant results per language (see examples in Appendix A).

3.4.1. Citation chaining

On completion of the selection of eligible reports identified through electronic searching, we searched the references sections of included papers for potentially eligible reports that had not been previously identified.

3.5. Selection process

Following deduplication, the first author screened titles and abstracts using Rayyan software (Ouzzani et al., 2016). Records clearly violating one or more inclusion criteria were excluded. The second author, blind to the first author's decisions, independently screened a randomly selected 10% sample ($n = 184$) of titles and abstracts. We compared decisions and discussed discrepancies ($n = 5$, $\kappa = 0.44$, moderate agreement) about which interventions met inclusion criteria until reaching agreement. Where an abstract did not explicitly violate inclusion criteria, full texts were sought. Where full texts were unobtainable online, we sought them through interlibrary loan or emailing authors. The first author screened 89 full texts, excluding any that violated inclusion criteria. A Korean applied linguistics colleague screened the six Korean papers. Ambiguous inclusion decisions not included in the prior collaborative 10% screening were discussed, and agreement reached.

3.6. Data collection process

Before completing final searches, we created a data extraction form (see Appendix B), adapted for relevance to the focus of our review from the principles laid out in the Cochrane Good Practice Guide (Cochrane Effective Practice and Organization of Care, 2017) and Boland, Cherry and Dickson (2017). We piloted the form on two included studies (Chou, 2014; Davis & Fan, 2016), ensuring it

Table 3
Search strategy.

(1) FL nature of studies	(2) Age and stage of participants and educational settings	(3) Nature of intervention	(4) Linguistic outcomes
ab (MFL OR EAL OR ESL OR EFL OR "foreign language*" OR FL OR "second language*" OR L2 OR French OR German OR Spanish OR English OR TEFL OR TESOL)	AND ab (KS1 OR KS2 OR KS3 OR KS4 OR "key stage" OR EYFS OR "early years" OR preschool OR kindergarten OR infant* OR junior* OR primary OR secondary OR elementary OR child* OR adolescent* OR "high school")	AND ab ("nursery rhyme*" OR choral OR chant* OR song* OR music* OR sing*)	AND ab (vocabulary OR gramma* OR phonolog* OR acquisition OR speaking OR spoken OR proficiency OR competence or skill*)

captured all relevant quantitative and qualitative data for extraction from PICOSS items (i.e., participants, intervention, comparator, outcomes, study design, setting; Boland et al., 2017), plus reference details and findings.

After the first author completed data extraction for 60 papers, the second author independently extracted data from 10% (n = 6) of the studies. To ensure a representative sample of full text reports and theses, two of the theses and four of the full reports were randomly selected. Any discrepancies were resolved through discussion.

3.7. Data items

The data items that were extracted from each report were as follows. Bibliographic information (authors' names, date of publication, publication source, full reference); language of publication; aims and research questions; design (this was inferred by the authors through careful reading of the methods sections of each report, and classified on the basis of the taxonomies provided by Campbell and Stanley (1963) and Shadish, Cook and Campbell (2002)); study duration, study location, school phase and socio-educational context (preschool, primary, secondary, public, private); description of the participants (age, gender, any information on special educational needs or further contextualising information); first and target languages; description of the singing intervention and comparator (if present); number of participants recruited and available for follow-up; group characteristics at baseline; outcome type (e.g., writing, reading, speaking, listening, vocabulary knowledge, fluency, comprehension, etc.); outcome measures (e.g., standardised instruments such as the PPVT (Peabody Picture Vocabulary Test; Dunn & Dunn, 1981) or researcher designed tests); descriptive reporting of outcomes (e.g., means and standard deviations); and analytic reporting of outcomes (e.g., effect size or *t*-statistic). Where information was unavailable or reported in such a way to be unclear, this was noted.

3.8. Study risk of bias assessment

Critical appraisal of included studies is a vital part of the systematic review process to determine how much confidence we can have in the findings of included studies. Since there are over 500 appraisal tools available, and a lack of clarity on how to choose and use them (Hong & Pluye, 2019), this section outlines the rationale for choosing the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018; Pluye, Gagnon, Griffiths & Johnson-Lafleur, 2009) for this systematic review.

It is a regrettable shortcoming in the field of evidence synthesis in language education that quality appraisal is rare (Chalmers, Brown & Koryakina, 2023). Perhaps because of this, a dedicated tool for this purpose in this field has yet to be developed. While the MMAT was originally designed by healthcare researchers, it nonetheless allows assessment of a variety of research designs and can be used in any discipline. This includes language education. We note, for example, that many protocols for systematic reviews in language education registered on IDESR have adopted the tool, and we have seen it used in a number of published reviews (e.g., Richter, 2021; Schulz, Hamilton, Wonnacott & Murphy, 2023; Willis, Neil, Mellick & Wasley, 2019).

The principles of methodological rigour (and in our case rigour relating specifically to establishing casual relationships) apply regardless of field (Isaacs & Chalmers, 2023). For example, adopting measures to minimise allocation bias (such as random allocation or statistical matching), recruiting sufficient numbers of participants to minimise statistical imprecision, and the validity and reliability of the tools used to measure outcomes are all general methodological principles that apply to all intervention research. The MMAT facilitates assessment of the extent to which these principles have been adhered to in any given report of research. Moreover, unlike other quality appraisal tools commonly used in education systematic reviews, such as Cochrane Risk of Bias 2 (Higgins et al., 2011) and Cochrane ROBINS-I (Sterne et al., 2016), which are design-specific, the MMAT allows for appraisal of a variety of designs under one coherent taxonomy.

The MMAT contains five methodology categories for assessing study quality across qualitative, randomised controlled trials, non-randomised comparisons, quantitative descriptive, and mixed methods studies. The MMAT has five criteria within each category which can be rated *Yes*, *No*, or *Can't tell*, with space for explanatory comments. The MMAT creators discourage giving a numerical score for each appraisal, instead encouraging reviewers to comment on how criteria were assessed and to justify those decisions in their reporting (Hong et al., 2018). Because the aim of this systematic review was to provide a comprehensive overview of the intervention research evaluating the use of songs with YLLs, low methodological or reporting quality was not considered a reason for exclusion. However, understanding the methodological quality of these studies helps researchers and practitioners understand the relative strength of the gathered evidence and thus informs policy and practice decisions, and signals areas where more research is needed.

This tool permits systematic (i.e., explicit, transparent, and replicable) application of study quality appraisal criteria across included studies. In this field, where unfalsified theoretical hypotheses often underpin confident proclamations about songs' effectiveness as YLL pedagogy, and this is reflected in teachers' beliefs about using songs (see Section 2), an objective and rigorous tool such as the MMAT helps to ensure review conclusions reflect the trustworthiness of included evidence.

The second author independently appraised six included studies (two randomly chosen theses, four full reports) after the first author completed the MMAT for all studies in the corpus. Interrater reliability was $\kappa = 0.56$, indicating only moderate agreement due to the potential for subjective interpretation of MMAT Q3.1 and Q3.5 in educational research contexts. Both authors discussed the interpretation of those items and resolved disagreements through discussion.

3.9. Synthesis methods

Where a body of literature includes diverse interventions and outcomes, as is often the case in social sciences research, it is inappropriate to conduct a meta-analysis of the studies' results (Petticrew & Roberts, 2006). Thus, following Petticrew and Roberts

(2006), we conducted a narrative synthesis as follows: (i) studies are grouped into comparable categories based on outcome measures; (ii) findings and quality appraisal of studies within each category are analysed; (iii) findings from all groups are synthesised narratively.

The groups for this synthesis arise from the reported outcome measures as follows: studies measuring vocabulary acquisition (splitting receptive and productive vocabulary measures into subgroups); studies measuring grammar outcomes (with verb and word-order studies as subgroups); studies measuring speaking skills (with pronunciation as subgroup); studies measuring listening skills; and studies measuring reading and writing skills. Studies that report outcome measures in sufficient detail are tabulated by category in Section 4.2.5 with reference to their measures, the claims made about the findings (e.g., whether there was a statistically significant effect of treatment on the outcome measures), and MMAT quality appraisal rating (strong, moderate, or limited confidence in the findings). Findings that support the hypothesis that songs aid language learning are coloured green; equivocal or mixed findings are coloured yellow; significant differences in favour of the control group are coloured pink. Combined with the MMAT colour-coding (green = strong, yellow = moderate, pink = limited trustworthiness ratings), it is possible to visualise positive or negative findings and the trustworthiness of studies within each category.

4. Results

4.1. Study selection

Fig. 1 shows the PRISMA flow diagram results of the selection and screening process. 2868 records were identified, including 1007 duplicates. Citation chaining identified one potentially eligible paper. Of the 94 full texts sought for retrieval, five were unavailable through interlibrary loan and contacting the authors proved unfruitful. They were, therefore, excluded. Three texts that, based on their abstracts, appeared to meet inclusion criteria were excluded because their participants were L1 learners (Crosswhite, 1996; Joyce, 2011; Lehman, 2019).

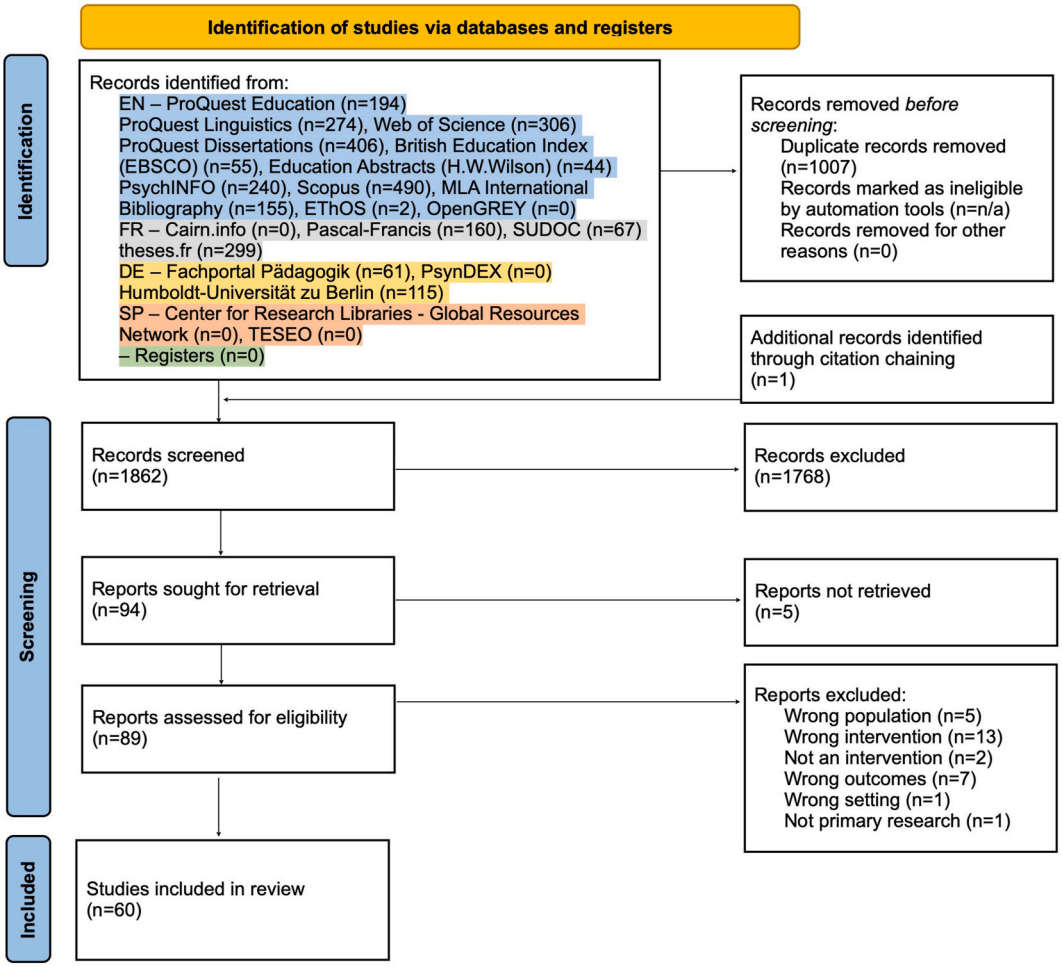


Fig. 1. PRISMA 2020 flow diagram of study selection process.

Table 4
Study characteristics
JA = journal article, PhD = doctoral thesis, MSc = master's thesis. Study duration (*in weeks unless stated otherwise).

Study	Publication status	Study design	Country	Sample size	Setting	Study duration (weeks*)	General outcomes	Specific outcome measures
1. Albaladejo, Coyle & Larios (2018)	JA	Single group pre/post	Spain	17	Preschool	6	Vocabulary	PPVT; observation of behaviour
2. Alinte (2013)	JA	Non-equivalent groups pre/post-test	Romania	34	Secondary	15	Grammar; attitudes	Grammatical knowledge test
3. Allen-Tamai (2000)	PhD	Non-equivalent groups pre/post-test	Japan	62	Preschool	11	Phonological awareness	Rhyme awareness
4. Alley (1988)	PhD	Non-equivalent groups pre/post-test	USA	47	Secondary	5	Listening; attitudes	Listening tests; attitudes to presentation mode
5. Al-Mosawi (2018)	JA	RCT	Iraq	40	Primary	12	Four skills	Four skills: reading, writing, listening and speaking
6. Amiri & Sobouti (2016)	JA	RCT	Iran	60	Preschool	8	Speaking	Pronunciation, fluency, grammar and vocabulary
7. An (2009)	JA	Non-equivalent groups pre/post-test	Korea	79	Primary	4	Vocabulary; attitudes	Vocabulary listening, comprehension of vocabulary meaning, speaking skills; attitude towards learning English
8. Au (2013)	JA	Cluster RCT	Hong Kong	126	Primary	18	Speaking	L2 or 2nd dialect accent
9. Augustine (2015)	JA	Non-equivalent groups pre/post-test	Malaysia	40	Preschool	6	Reading	Print knowledge, definitional vocabulary, phonological awareness
10. Becerra Vera & Luna (2013)	JA	Single group pre/post	Spain	49	Primary	1 school year	Listening	Listening tests
11. Boey (1978)	JA	Non-equivalent groups pre/post-test	Malaysia	573	Primary	2 school years	Four skills	Speaking, listening, reading, dictation
12. Busse, Hennies, Kreutz & Roden (2021)	JA	Non-equivalent groups pre/post-test	Germany	57	Primary	9	Vocabulary; grammar; attitudes	Vocabulary recall (name items); grammar translation; multiple choice grammaticality judgement task; affective outcomes of lessons
13. Caleya, Nieto & Espejo (2013)	JA	Non-equivalent groups pre/post-test	Spain	193	Primary	1 school year	Speaking	Pronunciation, accuracy, fluency, eagerness to repeat, accent, memorising
14. Campfield & Murphy (2013)	JA	RCT	Poland	87	Primary	3	Grammar	L2 word order; knowledge of function words
15. Chae & Yoon (2013)	JA	Non-equivalent groups pre/post-test	Korea	60	Primary	12	Memory; grammar; affective domains	Short/long-term memory (cloze tests); grammar; affective responses to input (story or song) and interest in learning English
16. Cheippe (2012)	PhD	Non-equivalent groups pre/post-test	France	20	Primary	7	Speaking	Pronunciation (L2 vowels)

(*continued on next page*)

Table 4 (*continued*)

Study	Publication status	Study design	Country	Sample size	Setting	Study duration (weeks*)	General outcomes	Specific outcome measures
17. Chen (2011)	PhD	Cluster RCT	Taiwan	128	Primary	12	Vocabulary; speaking; attitudes	Picture vocabulary test; phonemic analysis test; attitudes to music intervention
18. Chiang (2003)	PhD	Cluster RCT	Taiwan	120	Primary	18	Listening	Multiple choice listening comprehension & dictation
19. Chou (2014)	JA	Non-equivalent groups pre/post-test	Taiwan	72	Primary	5 × 100-min lessons	Vocabulary; attitudes	Written receptive vocabulary recognition (true/false, matching) and spelling/productive vocabulary writing (anagrams/gap-filling with pictures)
20. Coyle & Gómez Gracia (2014)	JA	Single group pre/post	Spain	25	Preschool	7	Vocabulary; attitudes	Receptive (picture recognition) and productive (naming task) vocabulary tests
21. Cruz-Cruz (2005)	PhD	Non-equivalent groups pre/post-test	USA	28	Primary	6	Vocabulary; grammar	Grammar (productive/judgement): pronouns, pronoun-verb agreement, adjectives, adverbs, articles; vocabulary: circle correct word to complete sentence; definition-word matching
22. Davis & Fan (2016)	JA	Single group pre/post	China	64	Preschool	7	Vocabulary; grammar; attitudes	MLU of productive description of picture card prompts
23. Diakou (2014)	PhD	Single group pre/post	Cyprus	171	Primary	2	Vocabulary; grammar	Pre-post questionnaires assessing participants' vocabulary/grammar attitudes; focus groups discussing acquisition; video observations tracing acquisition.
24. Dominguez (1991)	PhD	Non-equivalent groups, post-test only	USA	51	Primary	7	Reading	Basic reading skills (e.g., word recognition, digraphs, end sounds, letter sounds, referents, drawing conclusions, predicting outcomes, etc.)
25. Fonseca-Mora, Jara-Jiménez & Gómez-Domínguez (2015)	JA	Non-equivalent groups pre/post-test	Spain	63	Primary	11	Reading	Early grade reading assessment: letter name knowledge, oral reading fluency, initial sound identification
26. Good, Russo & Sullivan (2015)	JA	Cluster RCT	Ecuador	38	Primary	2 weeks (with follow-up test after 6 months)	Speaking; vocabulary	Pronunciation (vowel & consonant production); recall words/phrases from lyrics; translate English vocabulary into Spanish
27. Gorjian, Hayati & Barazandeh (2012)	JA	RCT	Iran	56	Primary	3 months	Vocabulary	Researcher designed vocabulary test with 14 items
28. Haghverdi (2015)	JA	RCT	Iran	60	Secondary	8	Listening; vocabulary/grammar;	Listening; vocabulary/grammar; reading (not defined further)

(*continued on next page*)

Table 4 (*continued*)

Study	Publication status	Study design	Country	Sample size	Setting	Study duration (weeks*)	General outcomes	Specific outcome measures
							reading; attitudes	
29. Hakozaki & Nakagawa (2020)	JA	Single group pre/post	Japan	91	Primary	6	Speaking	Pronunciation, overall intelligibility
30. Herrera, Lorenzo, Defior, Fernandez-Smith & Costa-Giomi (2011)	JA	Non-equivalent groups pre/post-test	Spain	97	Preschool	2	Phonological awareness	Phonetic awareness, verbal memory, naming speed, name and sound letters knowledge
31. Hsu (2009)	PhD	Non-equivalent groups pre/post-test	Taiwan	47	Preschool	6–8	Vocabulary; speaking	Pronunciation and oral spelling of colours
32. Jarvis (2013)	JA	Non-equivalent groups pre/post-test	UK	12	Primary	Not reported	Speaking; listening; attitudes	Speaking assessment of weekly target vocabulary; observation of behaviour; attitudes of staff to introducing MFL in EY setting
33. Jeong & Kim (2014)	JA	Non-equivalent groups pre/post-test	Korea	40	Primary	2 months	Listening; vocabulary; attitudes	Listening; vocabulary; attitudes to learning English
34. Kim & Kang (2015)	JA	Single group pre/post	Korea	128	Secondary	10 months	Listening; attitudes	National listening comprehension tests
35. Kim & Park (2012)	JA	Non-equivalent groups pre/post-test	Korea	87	Primary	3 months	Vocabulary	Vocabulary proficiency test
36. Klohs (1994)	PhD	RCT	USA	72	Secondary	4.5	Grammar; writing; attitudes	Verb tenses; written paragraph assessed for communicative skills; attitudes to mnemonic skills taught/perceived vs actual usage of mnemonics in the tests
37. LeBrun (2019)	PhD	Cluster RCT	USA	142	Secondary	15 lessons	Vocabulary; grammar; reading; listening; attitudes	Vocabulary: matching/cloze/multiple choice Grammar: cloze sentence to fill with correct verb conjugation Reading/listening comprehension
38. Legg (2009)	JA	RCT	UK	62	Secondary	1 h	Vocabulary	Translate English phrases containing passé composé/imperfect verbs into French equivalent; translate weekdays
39. Leśniewska & Pichette (2016)	JA	Single group post-test only	Canada	24	Preschool	4	Vocabulary	PPVT
40. Lowe (1995)	PhD	Non-equivalent groups pre/post-test	Canada	53	Primary	5 months	Vocabulary; grammar; reading; speaking; music skills	Vocabulary: cloze/matching; oral grammar (put words in correct order); reading: true/false, gap-filling; pronunciation; music skills – describe, create, perform
41. Ludke (2010)	PhD	Non-equivalent	UK	59	Secondary	4	Vocabulary; grammar; attitudes	Cloze test of song lyrics; translation French > English

(*continued on next page*)

Table 4 (*continued*)

Study	Publication status	Study design	Country	Sample size	Setting	Study duration (weeks*)	General outcomes	Specific outcome measures
		groups crossover						
42. Luo (2019)	JA	Single group pre/post	China	50	Secondary	3	Vocabulary; attitudes	Use target words in a sentence; Chinese > English word translation
43. Ma (2004)	JA	Single group pre/post	Korea	48	Preschool	4	Vocabulary; story recall	Picture vocabulary test: point (receptive) and label (productive); child prompted to complete sentences by reading/singing along with story
44. Madani & Nasrabadi (2016)	JA	Non-equivalent groups pre/post-test	Iran	112	Preschool	1 month	Vocabulary	Vocabulary learning/retention
45. Mamdouh (2017)	JA	Non-equivalent groups pre/post-test	Spain	19	Secondary	10	Listening	Listening comprehension
46. McCormack & Klopper (2016)	JA	Single group pre/post	Australia	5	Primary	6	Speaking	Graphic melodic contouring to measure oral fluency
47. McCormack, Klopper, Kitston & Westerveld (2018)	JA	Single group pre/post	Australia	6	Primary	8	Speaking	Pronunciation
48. Medina (1991)	PhD	RCT	USA	48	Primary	6	Vocabulary	Picture vocabulary test: circle item that matches the word read aloud
49. Moradi & Shahrokhi (2014)	JA	Non-equivalent groups pre/post-test	Iran	30	Primary	5	Speaking	Pronunciation, intonation, stress patterns
50. Muzammil & Andy (2019)	JA	Single group pre/post	Indonesia	31	Preschool	Not reported	Vocabulary; speaking; phrases	Receptive/productive vocabulary; phrases: matching
51. Navarro, Quiroga & Diaz (2018)	JA	Single group pre/post	Chile	25	Primary	5	Speaking	Pronunciation: words, phrases and sentences
52. Priester (2011)	MSc	Single group pre/post	USA	15	Preschool	5	Vocabulary	Oral productive task and journal pictures
53. Santos Jimenez, Gallegos Ruiz & Gomez Hermosa (2017)	JA	Cluster RCT	Peru	48	Primary	Not reported	Vocabulary	Measures unclear
54. Schunk (1999)	JA	RCT	USA	80	Primary	1–2	Vocabulary	PPVT
55. Siebring (2004)	MSc	Non-equivalent groups pre/post-test	Canada	53	Primary	2	Grammar	Fossilised errors tested orally – complete sentence/respond to question with correct form
56. Tomczak & Lew (2019)	JA	Non-equivalent groups pre/post-test	Poland	31	Secondary	3 per study (x2)	Vocabulary	Multi-word unit productive knowledge
57. Toscano-Fuentes & de Vega (2018)	JA	Single group pre/post	Spain	50	Primary	12	Reading	Timed (1 min) silent reading fluency & word identification/segmentation
58. Wang (2005)	MSc	Non-equivalent groups pre/post-test	China	133	Secondary	4.5 months	Grammar; attitudes	Formative grammar, summative grammar and listening comprehension tests
59. Yousefi (2014)	JA	RCT	Iran	60	Secondary	2 months and 11 days	Vocabulary	Provide L1 equivalent of English vocabulary item

(*continued on next page*)

Table 4 (*continued*)

Study	Publication status	Study design	Country	Sample size	Setting	Study duration (weeks*)	General outcomes	Specific outcome measures
60. Zhaku-Kondri (2014)	JA	Cluster RCT	Macedonia	57	Primary	8	Vocabulary; grammar; attitudes	Grammar (verb tenses) in pre/post-tests; vocabulary in post-test

4.2. Study characteristics

Table 4 provides characteristics for the 60 included studies. Patterns in the data are illustrated in the subsequent sections about study publication details, educational and geographic context, research design, and reported outcomes.

4.2.1. Publication details

Fig. 2 illustrates publication trends in this research area from the oldest paper (1978) to the most recent (2021). In the three decades from 1978 to 2008, 13 studies meeting the inclusion criteria were published, with none published from 1979 to 1987; since 2009, a further 47 eligible studies were published, 23 of these from 2013 to 2016. There are 43 peer-reviewed articles and 17 theses (n = 3 master's, n = 14 doctoral).

4.2.2. Geographic context

Studies included in this review were conducted in 23 countries (Fig. 3), from all continents except Africa. 83% (n = 50) studies are published in English, followed by Korean (5), Spanish (4), and French (1).

4.2.3. Instructional context

Fig. 4 illustrates the breakdown of participating settings. 57% of studies (n = 34) took place in primary schools, with the remaining 43% split equally between preschool (n = 13) and secondary (n = 13) contexts. 41% of the primary school studies (n = 14) were conducted from 2012 to 2014 (see Fig. 2).

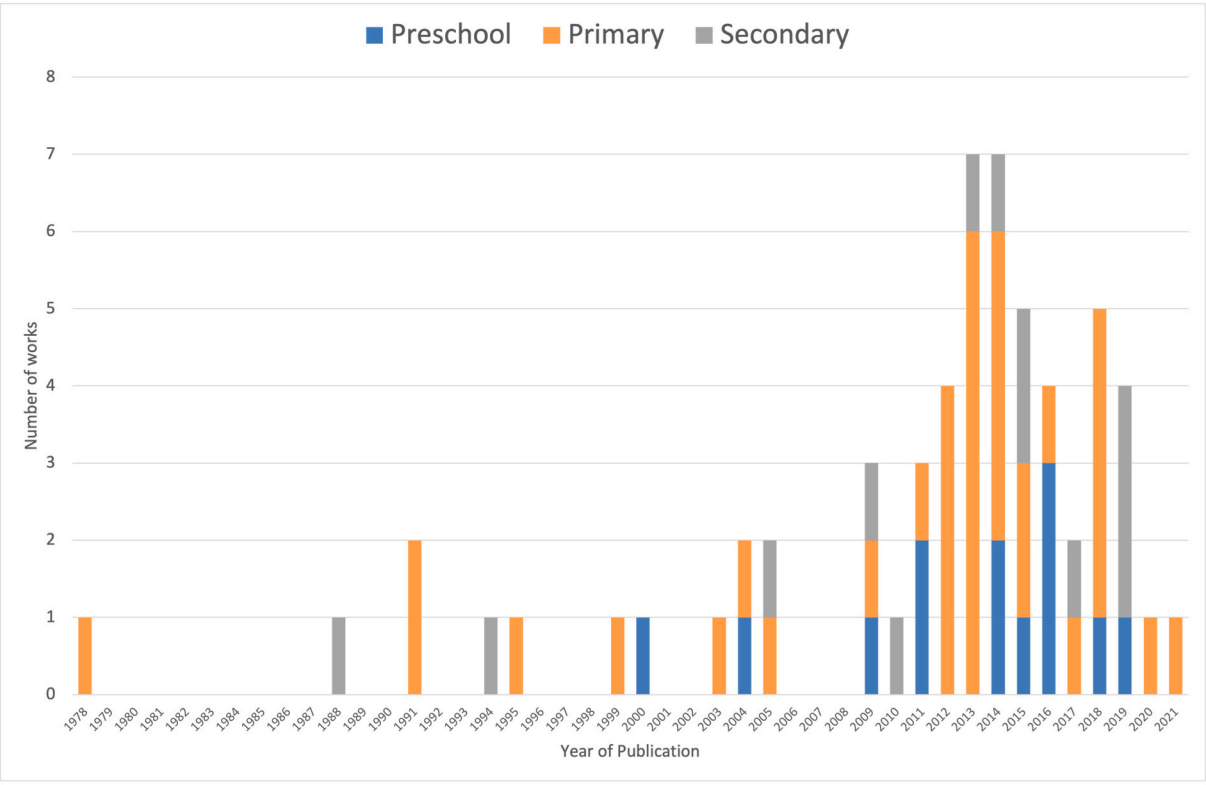


Fig. 2. Number of included studies by publication year and educational context.

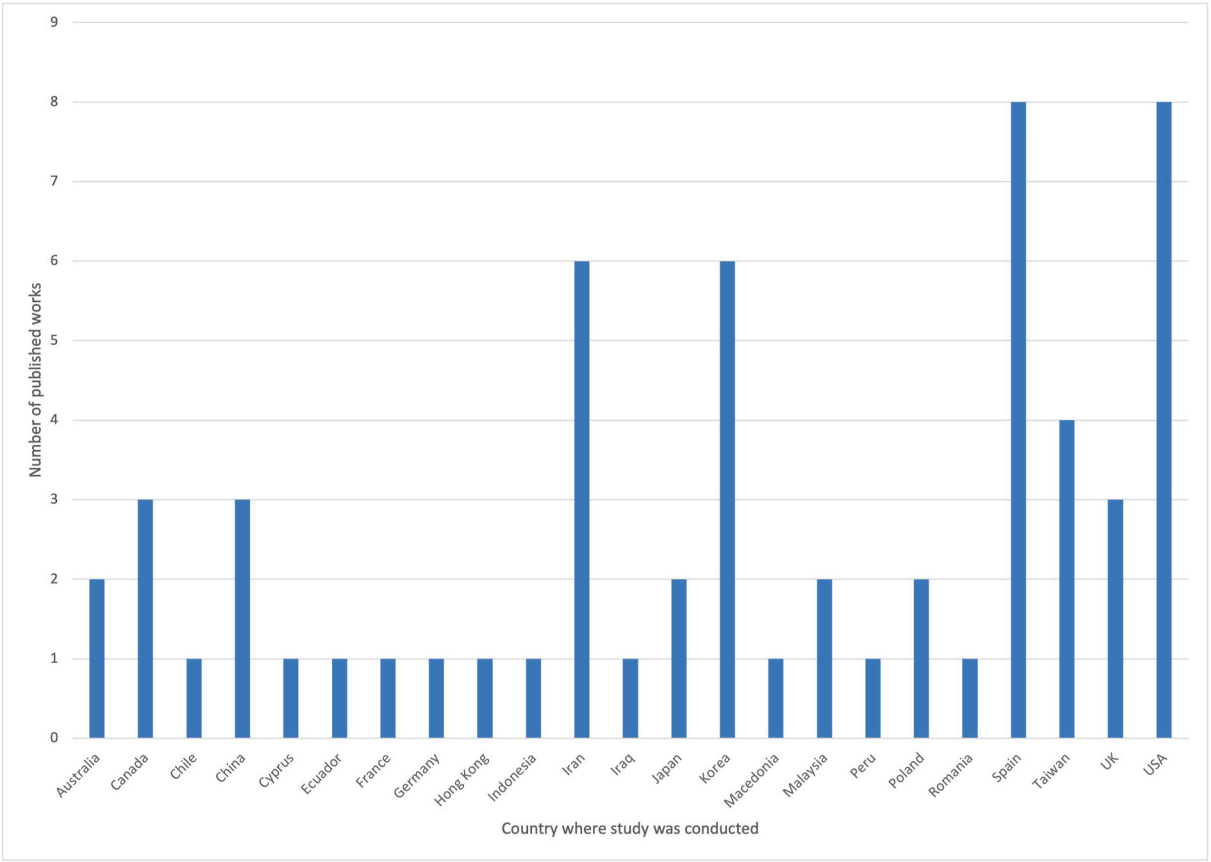


Fig. 3. Geographic region.

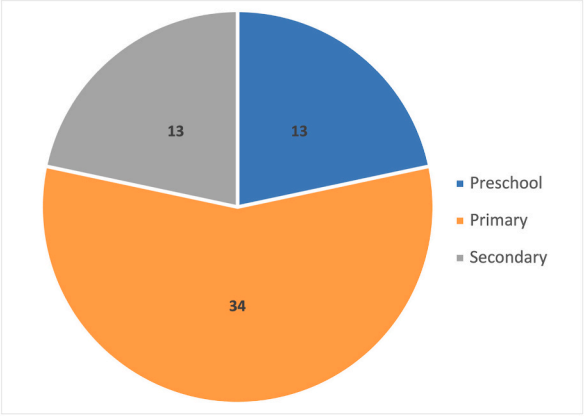


Fig. 4. Instructional context of studies.

4.2.4. *Study design*

Table 5 summarises included studies' designs.

4.2.4.1. Data type. All 60 studies used quantitative measures, with 48 reporting exclusively quantitative findings. 12 studies collected both qualitative and quantitative data[1,15,19,23,32,35,36,41,42,46,47,52].

4.2.4.2. Allocation strategy. Table 6 summarises how studies allocated participants to treatment or control conditions. Eight studies (13%) did not report any allocation strategy. 25 studies (42%) allocated intact classes, seven of which allocated classes randomly to

Table 5
Summary of study designs.

Study design	No. Of Works	Study ID[a]
Non-equivalent groups pre/post-test	25	2, 3, 4, 7, 9, 11, 12, 13, 15, 16, 19, 21, 25, 30, 31, 32, 33, 35, 40, 44, 45, 49, 55, 56, 58
Non-equivalent groups crossover	1	41
Single group pre/post-test	15	1, 10, 20, 22, 23, 29, 34, 42, 43, 46, 47, 50, 51, 52, 57
RCT	10	5, 6, 14, 27, 28, 36, 38, 48, 54, 59
Cluster RCT	7	8, 17, 18, 26, 37, 53, 60
Non-equivalent groups, post-test only	1	24
Single group, post-test only	1	39

[a] Superscript numbers refer to study ID in Table 4.

Table 6
Allocation strategy.

	Allocation strategy	No. Of works	Study ID
	Not reported/unclear from report	8	7, 16, 21, 32, 33, 35, 44, 49
CLUSTER	Intact classes (no strategy reported)	14	3, 9, 11, 12, 13, 15, 19, 25, 31, 40, 45, 55, 56, 58
	Intact classes (not randomly assigned)	4	2, 4, 41, 43
INDIVIDUAL	Random allocation of intact classes by drawing class names from a hat (first to be drawn assigned to music condition)	1	37
	Random allocation of intact classes (no strategy reported)	6	8, 17, 18, 26, 53, 60
	Random allocation at individual level (strategy not reported)	7	5, 6, 14, 27, 28, 38, 59
	Individuals matched by pre-test scores and randomly assigned to four groups, then groups assigned to conditions by shuffling papers with names of the groups on	1	48
	Matched by pre-test scores and randomly assigned to conditions by flipping a coin	1	36
	Matched by grade level, school and gender and assigned to conditions (allocation strategy not reported)	1	54
	Children's names alphabetised within their groups, assigned numbers, then odd numbers assigned to comparison and even to treatments (i.e., alternation)	1	24
	Stratified allocation by first language to four groups alternately (i.e., alternation)	1	30
	Allocation strategy not applicable as single group design	15	1, 10, 20, 22, 23, 29, 34, 42, 43, 46, 47, 50, 51, 52, 57

conditions with one reporting their allocation strategy. Twelve studies (20%) randomly allocated participants at an individual level, seven of which did not report their allocation strategy and four used different strategies. Fifteen studies (25%) used a single group pre/post-test design.

4.2.4.3. Study duration. Study duration ranged from one hour to two years. Three papers[32,50,53] fail to report duration. Fig. 5 illustrates the duration of remaining studies. Two studies[19,37] report how many lessons were taught and/or their duration, but not the period over which they were taught. 50% of the included studies lasted between two and nine weeks.

4.2.4.4. Control groups. 45 studies had control groups (or control items for within-subjects designs[1,22,29,39]), and 15 had none[8,10,13,19,20,23,34,43,46,47,50,51,52,55]. As Fig. 6 illustrates, of the 45 studies with a control group, two[5,53] did not report how the control was matched to the treatment group; four[26,27,28,45] generated control groups across multiple years in the same school; one group[54] was matched on level of English acquisition; one[36] was matched on the previous quarter's French exam grades; and 33 studies had a control group matched by age range[2,3,4,6,7,9,11,12,14,15,16,17,18,21,24,25,30,31,32,33,35,37,38, 40,41,42,44,48,49,56,58,59,60].

4.2.4.5. Sample size. Fig. 7 shows the sample size across included studies. Values ranged from 5 to 573, with a median of 56 participants. Removing two outliers with the smallest samples (five or six participants; half the number of the next largest sample) does not alter the median. Removing three outliers with 171, 193 and 573 participants alters the median to 53. Thus, neither extremely high nor low sample sizes affect the overall picture Fig. 7 presents.

4.2.5. General reported outcomes

Fig. 8 illustrates the outcomes reported by included studies: vocabulary, grammar, four skills (listening, reading, writing, speaking), attitudes, or other. Fig. 9 shows the total studies reporting each outcome type. These are not mutually exclusive: overall, 111 outcome assessments are reported. Over half of included studies (n = 33) used vocabulary measures, with 13 exclusively measuring vocabulary (over a fifth of total papers)[1,22,27,35,38,39,44,48,52,53,54,56,59].15 studies[2,6,12,14,15,21,23,28,36,37,40,41,55,58,60] included grammar measures, with two exclusively measuring grammar[14,55]. Five studies measured both vocabulary and grammar[12,21,23,41,60], and a further eight measured vocabulary, grammar plus another linguistic measure or attitudes[12,23,28,37,40,41,60].37 papers include measures

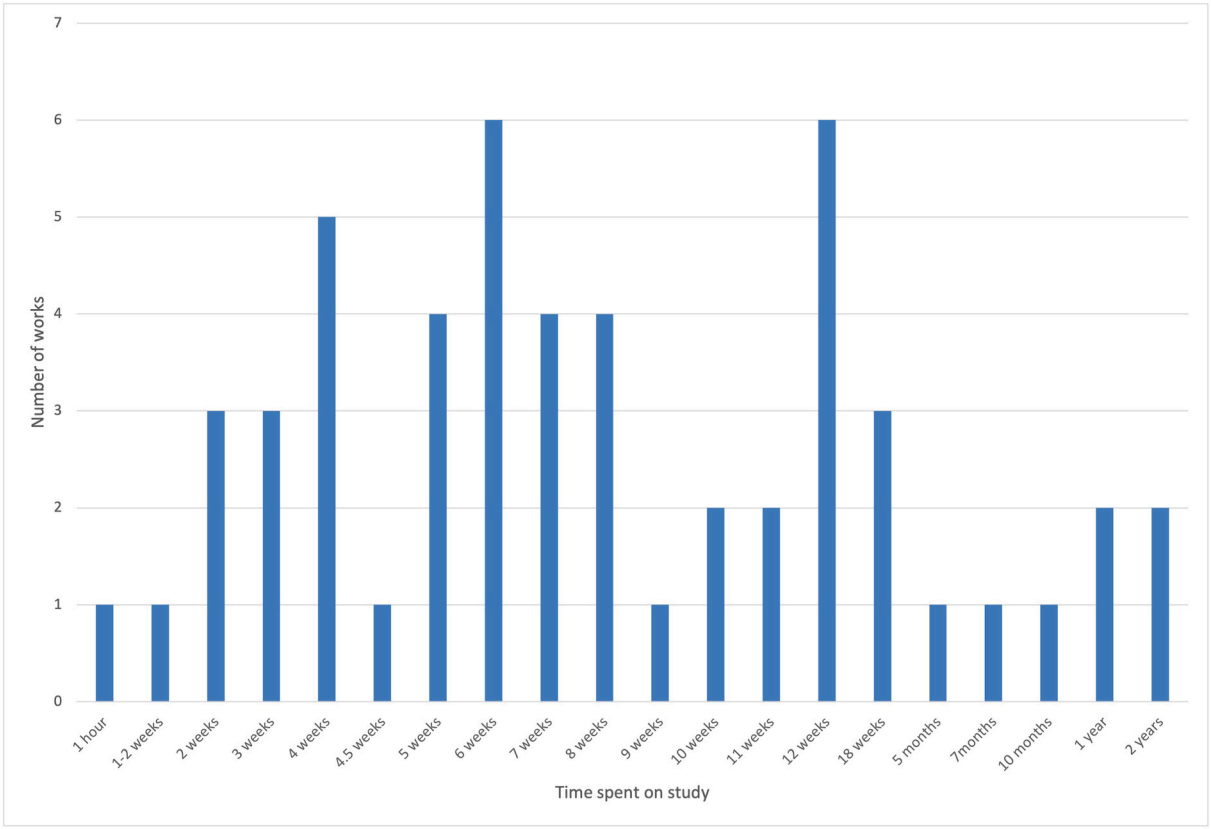


Fig. 5. Study duration.

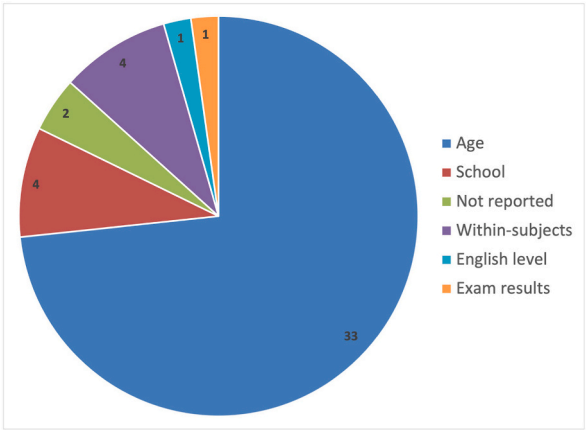


Fig. 6. How comparison groups are matched.

of the four skills: nine exclusively measure speaking skills[6,8,13,16,29,46,47,49,51], three listening skills[10,18,45], and three reading skills[9,24,25], while the remainder include a combination of skills and attitude measures[4,32,34]. Only two studies[5,11] measure all four skills. Two studies[3,30] measure phonological awareness. 18 studies included attitudinal measures [2,4,5,7,12,19,20,23,28,32,33,34,36,37,41,42,58,60]. Within these general outcomes, a variety of measures are reported. The following sections tabulate studies which reported their outcome measures in enough detail to permit further analysis.

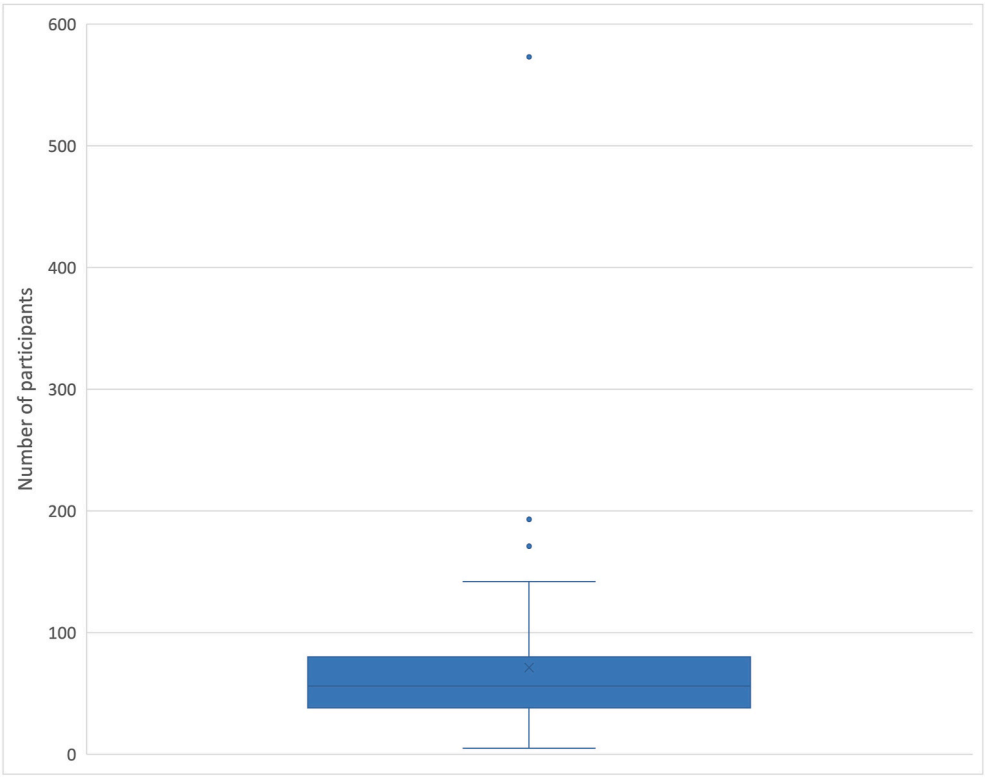


Fig. 7. Sample size.

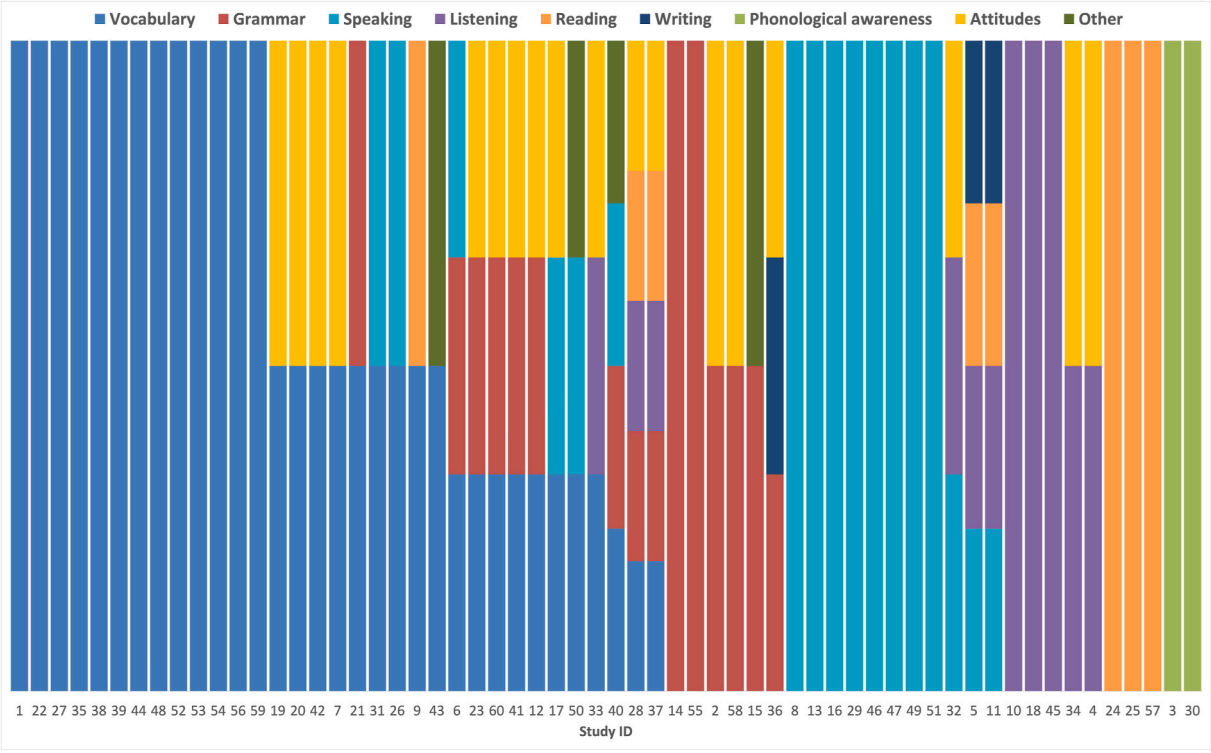


Fig. 8. Outcome type by study.

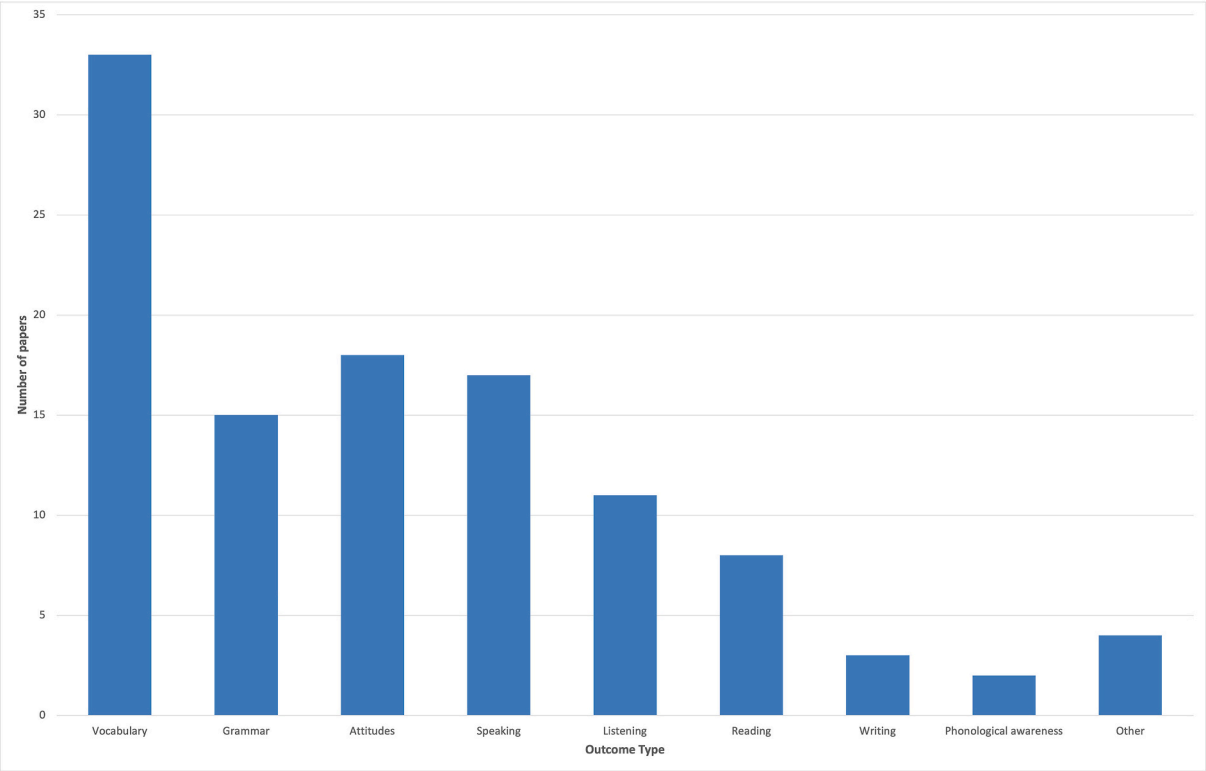


Fig. 9. Outcome frequency.

4.2.6. *Specific reported measures*

4.2.6.1. Vocabulary measures. The largest group of studies (n = 33) report vocabulary outcomes. 13 studies measure receptive vocabulary (summarised in Table 7) and 13 measure productive vocabulary (Table 8). Three studies measure both[19,20,43]. Ten studies[6,7,27,28,33,35,44,50,53,60] do not report clearly how vocabulary was measured.

Receptive vocabulary studies. 13 receptive vocabulary studies report seven types of receptive vocabulary measures. Seven studies used a picture vocabulary test, of which four reported using a standardised test, either the PPVT (Dunn & Dunn, 1981)[1,39,54] or TOLDP-3 (Newcomer & Hammill, 1997)[17]. They report equivocal results and received predominantly 'limited' trustworthiness ratings. Only one study with a 'strong' rating[17] reports a positive effect on the music treatment group's vocabulary scores from pre-to post-test. Yet one study cannot reliably claim for a universal positive effect of using songs on receptive vocabulary acquisition, especially when such a mixed picture arises from other studies. Any overall claims about songs' effectiveness for improving receptive vocabulary skills must be tempered with the knowledge that study designs are predominantly limited and outcome measures incomplete or incompletely reported. We simply do not know yet whether songs have any reliable effect that differs from other methods of presentation of new vocabulary.

Productive vocabulary studies. Table 8 summarises the 13 studies measuring productive vocabulary. Four papers[26,31,38,52] have no stated research questions, which on the MMAT indicates further appraisal may not be feasible or appropriate. Their findings should be treated cautiously.

Eleven studies claim singing has a positive effect on productive vocabulary. Davis and Fan's (2016)[22] comparison of singing or chanting conditions for vocabulary presentation to 'no presentation' has limited value here, since the question of import is 'Does presenting new vocabulary via song work better than alternative presentations?', not 'Does presenting words to children work better than not presenting words?' Two studies[41,56] involved aural input (songs or spoken conditions) with cloze practice exercises but only written output (cloze tests). Arguably changing modality from oral presentation to written production could influence students' performance on this kind of measure (see Murphy and Castillo's (2013) discussion of the implications of using one modality to teach and a different modality to assess).

One study[23] found songs have a positive effect on children's motivation to learn English as a foreign language, which the author claims in turn helps the children to learn more vocabulary from the songs. Since this paper ambiguously presents students' self-report in surveys and interviews about whether they feel they have learned vocabulary as a valid measure of progress, the findings contribute interesting and contextual but, ultimately, anecdotal evidence to the question of causal links between singing and vocabulary learning.

Two papers found equivocal effects of songs on productive vocabulary. The first[12] found no effect of songs on German EFL learners' productive vocabulary other than improving spelling, which the authors found surprising because the spoken condition group spent

Table 7
Studies reporting receptive vocabulary measures.

	Study ID	Receptive Vocabulary Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRESCHOOL	1. Albaladejo, Coyle & Larios (2018)	PPVT	Claim positive effect for songs but that songs alone performed worse than story or song/story combination.	No interpretive framework to guide qualitative findings; substantiating evidence provided briefly/descriptively in the paper; no real explanation of why mixed methods were used or integration of the qual/quant data.
	9. Augustine (2015)	Definitional vocabulary	Claim positive effect for songs for definitional vocabulary.	Baseline measures not reportedly fully (only p value); English L2 skills not reported clearly; No description of what happened in the intervention or whether it continued as intended.
	20. Coyle & Gracia (2014)	Receptive (picture recognition*)	Claims positive effect of songs on receptive vocabulary.	No control group; no report of which tests are used, or whether standardised measures; only prior knowledge of English is measured at baseline. Other confounders not accounted for. Potential test confounder as children took the same test 6 times.
	39. Leśniewska & Pichette (2014)	PPVT	Claim singing presentation condition worse than story or story/singing combined.	Only L1 receptive vocab measured as baseline and no other measures reported.
	43. Ma (2004)	Picture vocabulary test*: point (receptive) and label (productive); child prompted to complete sentences by reading/singing along with story.	Claim positive effect of singing on word recognition/labelling.	RQ2 seems inadequately addressed by the recall test results; non-standardised test of vocabulary recall; no baseline measures; insufficient information to know what happened during intervention.
PRIMARY	17. Chen (2011)	Picture vocabulary test from standardised Test of Language Development-Primary, 3rd Edition (TOLD:P-3: Newcomer & Hammill, 1997).	Positive effect claimed for vocabulary learning and pronunciation in song condition compared to control group with traditional methods.	Standardised tests (TOLD:P-3) used to test pronunciation and vocabulary; baseline measures of music and English experience showed no sig. differences between groups.
	19. Chou (2014)	Written receptive vocabulary recognition (true/false, matching).	Claims positive combined effect of songs, games and stories on vocabulary learning.	Unclear how data from observations were coded and interpreted, plus inadequate provision of data to support interpretation; no control group and songs/games/stories are mixed together in the intervention, so it is unclear what has any effect; self-assessment questionnaire unreliable way to measure vocabulary growth; vocabulary tests not in appendix and do not seem to tally up with target items (30 items, test with 25 marks).
	21. Cruz-Cruz (2005)	Circle correct word to complete sentence; definition-word matching.	Claims positive effect of songs on vocabulary and grammar.	RQs not stated (just aims/intention); researcher-designed non-standardised instruments to test grammar and vocab, with only % reported; only prior knowledge of English measured at baseline. Other confounders unaccounted for.
	40. Lowe (1995)	Vocabulary: cloze/matching.	Claims positive effect of music programme on composite French post-test but no Group X Time interaction effects for vocabulary alone.	Researcher-devised non-standardised measures of music, language, and maths achievement; no random allocation thus confounds inadequately accounted for.
	48. Medina (1991)	Picture vocabulary test: circle item that matches the word read aloud by tester.	Claims positive effect of music for low proficiency learners but since it is not significant at 0.5 level unclear why they claim this. No other positive effects. Very small samples of ¾ per group.	Researcher-devised, non-standardised oral vocabulary test where children circled the picture that corresponds with the word read aloud by the tester; only prior knowledge of English vocab is measured at baseline. Other confounders not accounted for at baseline.
	54. Schunk (1999)	PPVT	Claims positive effect of sung condition and spoken condition with signs compared to spoken text only. Sung/spoken with signs not significantly different to singing-only condition.	RQs not stated (just aims/intention); no baseline measures of cognition/control for L1 backgrounds.
SECONDARY	37. LeBrun (2019)	Vocabulary: matching/cloze/multiple choice.	Claims significant differences in vocab scores for junior high group for experimental group compared to control, but not when all groups added together.	Tests taken from the textbook, so not standardised measures; ANCOVA used to control for vocab, grammar, listening and reading baselines, but other baseline measures not reported and groups were intact classes.
	59. Yousefi (2014)	Provide L1 equivalent of English vocabulary item.	Claims positive effect of music on short and long-term retention of vocabulary.	RQs not stated (just aims/intention); data are insufficiently reported to know what they did/gathered; non-standardised measures; no baseline measures reported (pretest results unreported); unclear if groups are from the same or different schools.
		*unclear whether standardised PPVT was used	Green = positive, Yellow = mixed, Pink = negative	Green = strong, Yellow = moderate, Pink = limited

more time reading the words than the singing group. The second study[20] claims that songs positively affected 25 5–6-year-olds' receptive but not productive vocabulary. However, with no control group, no clear report of which measures are used, and limited account of confounding factors, little weight can be given to these findings as evidence of causality.

In summary, whilst 23 studies investigate songs' effect on sufficiently well described receptive and productive measures of vocabulary, the scope of the research to permit an overall analysis of effectiveness is limited by lack of rigorous and reliable design, data collection and reporting. Most authors claim to have found positive effects for singing on vocabulary measures, but only two papers[12,17] received strong trustworthiness ratings, one using receptive and one productive vocabulary measures. Overall, evidence is not substantial or reliable enough to make any strong causal inferences about the effect of singing on vocabulary uptake.

Table 8
Studies reporting productive vocabulary measures.

	Study ID	Productive Vocabulary Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRESCHOOL	20. Coyle & Gracia (2014)	Productive naming task.	Claim non-significant effect of singing on productive vocabulary.	No control group; no report of which tests are used, or whether standardised measures; only prior knowledge of English is measured at baseline. Other confounders not accounted for. Potential test confounder as children took the same test 6 times.
	22. Davis & Fan (2016)	MLU of productive description of picture card prompts.	Claim singing/chanting equally effective compared to no presentation control.	Unclear whether standardised test: measured MLU for productive output as cued by picture cards; potential confounders not reported. English level given holistically as "similar" to Grade 1/2 children. No differences between classes reported at all.
	31. Hsu (2009)	Pronunciation and oral spelling of colours.	Claims positive effect of singing on oral vocabulary and spelling of target words.	RQs not stated (just aims/intention); researcher-devised, non-standardised tests of oral vocabulary pronunciation (confound: recall and pronunciation simultaneously?) and oral spelling based on preLAS 2000; only prior knowledge of English is measured at baseline. Other confounders not accounted for at baseline.
	43. Ma (2004)	Picture vocabulary test*: point (receptive) and label (productive); child prompted to complete sentences by reading/singing along with story.	Claims positive effect of singing on oral target word & phrase recall.	RQ2 seems inadequately addressed by the recall test results; non-standardised test of vocabulary recall; No baseline measures; insufficient information to know what happened during intervention.
	52. Priester (2011)	Oral productive task and journal pictures.	Claims positive effect of singing on oral vocabulary and use of target words when drawing in journals.	RQs not stated (just aims/intention); non-standardised oral test of vocabulary, plus tally of researcher's observations and journal pictures; no baseline measures and no control.
	12. Busse, Hennies, Kreutz & Roden (2021)	Vocabulary recall (name items).	No effect of singing claimed, except for spelling.	Non-standardised, researcher-devised tests of written vocabulary, translation, and multiple-choice grammar.
PRIMARY	19. Chou (2014)	Spelling/productive vocabulary: writing (anagrams/gap-filling with pictures).	Claims positive combined effect of songs, games and stories on vocabulary learning.	Unclear how data from observations were coded and interpreted, plus inadequate provision of data to support interpretation; no control group and songs/games/stories are mixed together in the intervention, so it is unclear what has any effect; self-assessment questionnaire unreliable way to measure vocabulary growth; vocabulary tests not in appendix and do not seem to tally up with target items (30 items, test with 25 marks).
	23. Diakou (2014)	Pre/post questionnaires assessing participants' vocabulary ; focus groups discussing acquisition ; video observations tracing acquisition.	Claims positive effect of introducing songs on pupil interest/motivation, which in turn has positive effect on vocabulary uptake.	Questionnaire/self-report data are inappropriate measures of the linguistic outcomes included in the study; some children had English lessons outside school, there were mixed abilities, and self-report at pretest unreliable baseline; no comparison group for intervention.
	26. Good, Russo & Sullivan (2015)	Pronunciation (vowel & consonant production); recall words/phrases from lyrics; translate English vocabulary into Spanish.	Claim positive effect of songs on vocabulary recall.	RQs not stated (just aims/intention); not reported which tests are used exactly, or whether they are standardised measures; no pretest of pronunciation; demographic survey not reported; no baseline measures reported (e.g., cognition, musical aptitude).
SECONDARY	38. Legg (2009)	Translate English phrases containing passé 42erano42/imperfect verbs into French equivalent; translate weekdays.	Claims positive effect of music condition on learning song words & Eng>Fre translation.	RQs not stated (just aims/intention); non-standardised translation task; participants from same age/level and randomised into conditions with a spreadsheet, but no other baseline measures reported.
	41. Ludke (2010)	Cloze test of song lyrics; translation French > English.	Claims positive effect of singing on French language skills compared to visual art/drama.	Researcher-devised written cloze and grammar tests, which are in a different modality to the input (listening/reading) and exclude speaking; attrition means that data is collected for 75% of participants (n=16 missing data out of n=57 participants) which is below MMAT acceptable level - 80%; no cognition baseline.
	42. Luo (2019)	Use target words in a sentence; Chinese > English word translation.	Claims positive effect of singing on vocabulary learning.	No description of interview methods; insufficient report of findings or how they derive from interview data; researcher-devised productive (written) vocabulary and translation tests; non-standardised; baseline measures or group differences not reported; insufficient information to know what happened during intervention; no explanation of why mixed methods were chosen.
	56. Tomczak & Lew (2019)	Multi-word unit productive knowledge.	Claims positive effect of songs for learning MWU.	Researcher-devised gap-filling MWU exercise; researcher-designed background questionnaire results not reported hence confounders not adequately accounted for.
			Green = positive, Yellow = mixed, Pink = negative	Green = strong, Yellow = moderate, Pink = limited

4.2.6.2. Grammar measures. Table 9 summarises 12 studies measuring an aspect of grammatical learning with adequately reported measures, although three[6,58,60] did not report clearly enough to allow discussion of their findings. Three further studies[2,15,28] did not report their measures.

All four secondary and three primary studies focused on verbs. Overall, their findings are inconclusive about songs' influence on verb learning since none of their methodologies or participant demographics overlap enough for comparison. Two studies[14,40] variously investigated how songs influence their participants' learning of FL word order.

There is some trustworthy evidence from studies[12,14] that measure YLLs' grammatical learning yet since these studies focus on different aspects of grammar, few conclusions can be drawn beyond the studies themselves. Notably, Busse et al. (2021)[12] included six items in their multiple-choice verb test, whereas Campfield and Murphy (2013)[14] had 70 items in their GJT. The latter is arguably more reliable since it tests participants more robustly by requiring them to transfer learning from one context (treatment condition) to another (GJT). Most importantly, as well as a larger sample size, Campfield and Murphy (2013) used random allocation at the individual level. Thus, the associated increase in statistical power means it is less prone to false-positive results than studies allocating

Table 9
Studies reporting grammar measures.

	Study ID	Grammar Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRESCHOOL	6. Amiri & Soubouti (2016)	Combined pronunciation, fluency, grammar and vocabulary in 'YLE' (Young Learner English) test.	Claim large effect of singing on grammar test.	Variables not clearly defined and the instrument is not included in the report; No pretest of prior English knowledge ("their English background knowledge was almost the same"); participants encouraged to listen to materials outside the intervention time; Very little description of the intervention itself, so difficult to ascertain whether exposure occurred as intended. Learners in Exp group encouraged to review materials at home, and this variation is not accounted for.
PRIMARY	12. Busse, Hennies, Kreutz & Roden (2021)	6 question-answer pairs presented in English (3 from songs, 3 new): participants choose correct form of verb 'to do' in multiple choice.	Claim students in the singing group identified correct form of verb "to do" better than speaking/control group when sentences were already provided, with progress retained over retention period.	Non-standardised, researcher-devised tests of written vocabulary, translation, and multiple-choice grammar.
	14. Campfield & Murphy (2013)	L2 word order (70 sentences) and knowledge of function words (64 sentence pairs) tested with grammaticality judgement tasks.	Claim significant effect of song input on GJT for word order (particularly verb-last structures). No effect detected for function-words.	Clear report of measures and outcome data. GJTs are researcher-designed. Baseline measures of age, gender, mother's education, exposure to English, cognitive abilities, PPVT and grammar measures all showed no significant differences between groups.
	21. Cruz-Cruz (2005)	Grammar (10 questions in 6 sections) included productive (choosing the correct pronoun), judgement task (which agreement is correct?), cloze with articles provided to fill in a/an, 'spot the adjective' sentence, knowing if an adverb is of time or manner.	Claim experimental group outperforms control on grammar post test.	RQs not stated (just aims/intention); researcher-designed non-standardised instruments to test grammar and vocab, with only % reported; only prior knowledge of English measured at baseline. Other confounders unaccounted for.
	23. Diakou (2014)	Questionnaire and focus group questions about how songs help them learn grammar.	Claims songs helped students memorise grammar structures.	Questionnaire/self-report data are inappropriate measures of the linguistic outcomes included in the study; some children had English lessons outside school, there were mixed abilities, and self-report at pretest unreliable baseline; no comparison group for intervention.
	40. Lowe (1995)	Oral grammar: students asked to rearrange words to form a sentence (5 items) and read it aloud. Words in the incorrect order lost a mark.	Claims significant difference in favour of treatment group for oral grammar post-test, when achievement in French and maths taken as covariates.	Researcher-devised non-standardised measures of music, language, and maths achievement; no random allocation thus confounds inadequately accounted for.
	55. Siebring (2004)	Oral interviews targeting fossilised verb error structures.	No significant effect detected of treatment on improving fossilised verb errors.	Two versions of RQs reported; used Harvey (2004) guide to error correction in French – non-standardised test; no baseline measures other than pretest; CD of songs given to students to listen to at home, so confound of exposure; tester gave "think of the song" prompt when children did not answer, but does not report how often this happened.
	60. Zhaku-Kondri (2014)	Target verb tenses – I would/Would I?/I wouldn't – but unclear how exactly these are tested.	Claims significant effect of using song lyrics on grammar test score, helping pupils practise the grammar and understand spoken and written English.	RQs not stated (just aims/intention); vocabulary only measured in the post-test; no report of what the measures entailed or clear description of intervention; unclear whether data complete since the numbers in the groups differ in the paper at various points it is n=57, or n=60 in the results; no baseline measures other than pretest of grammar.
SECONDARY	36. Klohs (1994)	Change French sentences into past tense, then write justification in English of chosen tense.	Claims significant effect of mnemonic strategies on learning grammar.	Researcher-devised grammar and essay tasks; participants from three classes were matched before being randomly assigned to treatment groups but no other baselines are reported. Good integration of mixed methods data to draw conclusions.
	37. LeBrun (2019)	Cloze sentences: fill in blank with correct form of verb. Write a response to the question in Spanish.	No effect detected in singing condition. Significant difference in favour of control group (total participants – all ages added together).	Tests taken from the textbook, so not standardised measures; ANCOVA used to control for vocab, grammar, listening and reading baselines, but other baseline measures not reported and groups were intact classes.
	41. Ludke (2010)	Translate 5 sentences Fre>Eng from song and 5 from dialogue with "acceptable" scores used as basis for statistical analysis when Eng meaning was close to correct Fre meaning (e.g., only one incorrect verb tense or form).	Both age groups improved grammar from pre- to mid-point test, but only older age group improved from mid- to post-test (and younger group's score decreased).	Researcher-devised written cloze and grammar tests, which are in a different modality to the input (listening/reading) and exclude speaking; attrition means that data is collected for 75% of participants (n=16 missing data out of n=57 participants) which is below MMAT acceptable level – 80%; no cognition baseline.
	58. Wang (2005)	Form-changing and picture-writing test of 3 English verb tenses (unclear what this means in practice).	Claim experimental group is more competent in using target grammatical rules, as shown by them scoring significantly higher on form-changing and picture-writing (but not multiple choice) tests.	Intervention and data are insufficiently reported to know what they did/gathered or to evaluate any confounding factors as groups inadequately described.
			Green = positive, Yellow = mixed, Pink = negative	Green = strong, Yellow = moderate, Pink = limited

intact classes where the sample is n = 2.

4.2.6.3. Speaking measures. 17 studies measure L2 speaking skills, five[5,11,13,32,50] not clearly reporting how outcomes were measured. Table 10 summarises findings from the remaining 12 studies. Ten report using pronunciation measures, with seven[8,29,40,46,47,49,51] investigating the effect of song treatment conditions on participants' accent or intelligibility at word, phrase or sentence level, two[16,26] investigating the effect of songs on pronunciation at the level of vowel and/or consonant sounds, and one[17] investigating pronunciation of phonemes with or without music treatment. All the studies bar one[40] report positive effects of music treatment on the various

Table 10
Studies reporting speaking measures.

	Study ID	Speaking Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRESCHOOL	6. Amiri & Soubouti (2016)	Combined pronunciation, fluency, grammar and vocabulary in 'YLE' (Young Learner English) test.	Claim that all four subskills of speaking (pronunciation, fluency, grammar and vocabulary) were statistically significantly improved in the song group, compared to the control.	Variables not clearly defined and the instrument is not included in the report; no pretest of prior English knowledge ("their English background knowledge was almost the same"); participants encouraged to listen to materials outside the intervention time; very little description of the intervention itself, so difficult to ascertain whether exposure occurred as intended. Learners in Exp group encouraged to review materials at home, and this variation is not accounted for.
	31. Hsu (2009)	Oral test of colours (can the child recall and pronounce the colour that corresponds to the colour card and "what colour is this?" question) and give the oral spelling. 1 point for correct pronunciation, 1 point for correct spelling.	Claims rhythmic teaching methods help EFL kindergarteners acquire target vocabulary pronunciation and spelling.	RQs not stated (just aims/intention); researcher-devised, non-standardised tests of oral vocabulary pronunciation (confound: recall and pronunciation simultaneously?) and oral spelling based on preLAS 2000; only prior knowledge of English is measured at baseline. Other confounders not accounted for at baseline.
PRIMARY	8. Au (2013)	Participants read two illustrated stories aloud (one English, one Putonghua) after hearing NS of each language read story aloud. Accents rated on five-point scale by three NS of each language.	Claim significant positive effect on pronunciation of ambient Putonghua music on Cantonese L1 second-dialect learners of Putonghua. No measurable benefits detected for English songs on L2 pronunciation not closely related to L1.	Potential confound in accent rating scores. There are different stories in the Chinese/English tests: this could be a confounding factor as one is about sport and the other about animals with much more repeated vocabulary.
	16. Chieppe (2012)	Participants read text aloud and recordings are transcribed, with target German vowels/diphthongs rated by NS for NS norm pronunciation.	Claims improvement of singing groups on target German vowel and diphthong sounds.	RQs not clearly set out but scattered questions from pp.7–15 explain the aim of investigating the effect of songs on pronunciation; researcher-designed pronunciation measures; lack of comprehensive baseline data (e.g., cognition) but abilities split across four groups according to reading level.
	17. Chen (2011)	Phonemic analysis test from TOLD-P: 3. 14 items measured children's pronunciation of phonemes and their ability to break down spoken words into shorter phonemic portions.	Claims students' pronunciation gain scores were statistically and significantly affected by music treatment, even when taking current private music lessons into account as a covariate.	Standardised tests (TOLD:P-3) used to test pronunciation and vocabulary; baseline measures of music and English experience showed no sig. differences between groups.
	26. Good, Russo & Sullivan (2015)	Pronunciation of vowels tested with support of lyrics handout: children asked to reproduce the lyrics (not specified whether to sing or speak them). 15 target vowels/consonants rated 1 for correct pronunciation (i.e. English not Spanish norms).	Claim sung condition better than spoken condition for teaching vowel sounds, but no significant difference in pronunciation of consonant sounds.	RQs not stated (just aims/intention); not reported which tests are used exactly, or whether they are standardised measures; no pretest of pronunciation; demographic survey not reported; no baseline measures reported (e.g., cognition, musical aptitude).
	29. Hakozaki & Nakagawa (2020)	Participants read a familiar text aloud. Segmental features, sentence level stress, and overall intelligibility all scored on a scale of 1–5 (1 = poor, 5 = high) by three native English speakers.	Claims that chants had a significant effect on intelligibility of English pronunciation by helping Japanese EFL learners focus on prosodic features of English.	Two texts read aloud (92 recordings) and evaluated by 3 judges with Cronbach's alpha range 0.74–0.90, but unclear where the scale comes from and if it is standardised. No baseline imbalances are reported, and tests are non-standardised.
	40. Lowe (1995)	Read five sentences aloud. Pronunciation scored on a five-point scale by five French immersion teachers. Average pre- and post-test scores for each student are used in analyses.	No effect of music condition found for pronunciation measure alone, but overall composite French score was significantly different for treatment group.	Researcher-devised non-standardised measures of music, language, and maths achievement; no random allocation thus confounds inadequately accounted for.
	46. McCormack & Klopper (2016)	L2 oracy progress measured with graphic contouring (visual representation) of pronunciation of a marker sentence once a week for six weeks.	Claim increased oracy and fluency in all six students.	RQs not stated (just aims/intention); no comparison group; unclear whether music program or repeated testing of single sentence responsible for increased speed of elicited speech samples.
	47. McCormack, Klopper & Westerveld (2018)	Weekly speech samples collected and analysed using the Student Oral Language Observation Matrix [SOLOM] (California Department of Education, 1981), and the EAL/D Rating Scales designed by the research team.	5/6 EAL/D participants' English pronunciation improved, 1 decreased according to both measures. Although students' native accent was retained, their speech was more coherent post-intervention in comparison to their pre-intervention results.	RQs not stated (just aims/intention); researcher-designed measures; no control group; baseline of pronunciation but no other baseline measures.
	49. Moradi & Shahrohki (2014)	Post-test of pronunciation, intonation, stress recognition (each marked out of 10). Recordings of post-test compared with original song input pronunciation.	Claim positive effect of treatment on pronunciation (segmental), and intonation and stress recognition (suprasegmental articulation).	Insufficient information to know what happened during intervention or exactly what it entailed; non-standardised test of pronunciation; no baseline measures other than textbook levels test.
	51. Navarro, Quiroga & Diaz (2018)	English pronunciation evaluated at the level of words, phrases, and sentences. Repeat words after hearing recording (1); choose three objects and describe them (2); do an oral presentation (3). Marked according to whether Adequate, sufficient, or insufficient for 1 & 2; or Excellent, good, sufficient, and insufficient (3).	Claim positive effect of treatment on students' pronunciation. None remained in 'insufficient' grading after interventions.	RQs not stated (just aims/intention); intervention and data are insufficiently reported to know what they did/gathered; no baseline measures and group differences not reported.
			Green = positive, Yellow = mixed, Pink = negative	Green = strong, Yellow = moderate, Pink = limited

pronunciation measures. Only one paper[17] received a 'strong' trustworthiness rating, hence these findings present a questionable picture of the effect of song instruction or ambient input[8] on students' L2 pronunciation. One paper lacks any inferential statistical analysis[51], instead reporting percentage increases in each score band, and thus cannot reliably detect a treatment effect. Six lacked clearly defined research questions[16,26,31,46,47,51] making it impossible to assess the precise aim of the research and therefore the relationship of the findings to those aims. Their findings should be interpreted cautiously as evidence of songs' influence on L2 pronunciation.

To summarise, the largest group of studies measuring speaking skills focused on pronunciation measures, albeit at levels from single sounds to whole sentences and using different measurement tools. A positive effect of singing on speaking outcomes was claimed by all but one paper. The predominantly limited trustworthiness ratings for most studies should be considered when evaluating the evidence in this area.

4.2.6.4. Listening measures. 11 included studies investigate the effect of singing interventions on L2 listening skills but

Table 11
Studies reporting listening measures.

	Study ID	Listening Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
SECONDARY	4. Alley (1988)	Weekly unit test where text was spoken/sung to match treatment conditions. Comprehensive end-of-treatment exam testing all content, through narrative or dialogue only (no sung presentation). No report of the actual test content.	No significant differences between either song or listening skills (active control) treatment groups on weekly unit tests or post-test. Treatment groups scored significantly higher than no treatment groups in post-test. Inconclusive: both treatment groups did better than groups with no focus on listening skills.	RQs not stated (just aims/hypotheses); researcher-designed unit tests and post-test differed (text presented as song in units and not in post-test, narrative only) and no appendices to check questions, so this is unclear. Exp group n=24 on pretest, and fluctuates on unit tests from n=14-21, so variable attrition down to 66% (potentially incomplete data if <80%?). No baselines reported other than pretest.
	37. LeBrun (2019)	Test from the textbook/course. Listening comprehension of 3-minute Spanish audio recording with 10 yes/no questions to check understanding.	No significant differences detected between treatment and control groups (composite group or within age categories).	Tests taken from the textbook, so not standardised measures; ANCOVA used to control for vocab, grammar, listening and reading baselines, but other baseline measures not reported and groups were intact classes.

nine[5,10,11,18,28,32,33,34,45] fail to report their measures in enough detail to synthesise. Table 11 summarises findings from two studies[4,37] that report listening measures more substantially. Neither found a statistically significantly different effect of music treatment on listening skills compared to alternative treatment groups, but Alley (1988)[4] found that both treatment and comparator groups outperformed classes who received no treatment.

These studies have several methodological limitations. Alley (1988) reports variable attrition rates for all end-of-unit tests (down to 66% at times), which means the data is incomplete (following Petticrew and Roberts' (2006) benchmark of no more than 20% attrition rates). Both studies fail to report baseline measures other than the pre-tests, thus cognitive ability and other confounders are unaccounted for in these quasi-experimental designs. Neither study used standardised tests to measure listening outcomes. Overall, there is little existing evidence for whether singing-based music interventions have a demonstrable effect on acquisition of L2 listening skills.

4.2.6.5. Reading measures. Eleven studies investigate L2 reading outcomes, including two[37,40] with measures of reading comprehension and five[3,9,24,25,30] with measures of reading skills components such as phonological awareness, naming speed or sound identification. One study[57] measured reading fluency. Table 12 summarises these eight studies. Three others[5,11,28] did not report their reading measures.

Two reading comprehension studies[37,40] report contradictory findings for the influence of singing treatment on reading skills. Both studies allocated intact classes rather than randomising individuals to experimental conditions, thus systematic differences between groups (biases) cannot be ruled out as explaining the differences. They cannot be statistically synthesised effectively due to differences in educational context (primary foreign language and secondary immersion settings), participants' ages, and diverse methodology.

It was challenging to draw conclusions about individual or overall findings from papers reporting phonological and other reading skills component measures. One study[25] has no clear research questions, resulting in a limited trustworthiness rating despite meticulous reporting. Additionally, their non-musical treatment group comprised Spanish gypsy families' children, who may have a strong sense of rhythm from increased childhood exposure to music (Gil & Azcune, 2012), a potentially confounding factor that is neither controlled for nor reported until the discussion. Another paper with several unaccounted confounding factors[30] found that L2 Spanish learners benefitted most from the phonological training with music condition, which has interesting implications for SLA contexts. Research questions were not stated clearly, and this was a two-year study with two eight-week intervention periods. Cognitive measures were taken at the beginning, but cognitive ability in such young learners (aged 4–5 years) could change over two years, affecting the findings' reliability. Both papers report that phonological training with and without music improves performance on a range of reading assessments.

The reading fluency study[57] used subtitled music videos to support Spanish L2 English learners (age 9–10 years) with phoneme-graphene correspondences and decoding skills during timed silent reading tests. It reported positive findings but was a single-group pre/post-test design with no non-music comparison, and reported descriptive frequency statistics of participants' outcomes. Whilst these papers provide some promising avenues for future investigation, no reliable conclusions can be drawn about songs' influence on learners' reading outcomes.

4.2.6.6. Writing measures. Three included papers[5,11,36] report measuring writing outcomes. Al-Mosawi (2018)[5] reports no test content details, making it unclear how using nursery rhymes on YouTube substantially increased pupils' writing development. Table 13 summarises two remaining studies. Since they found no positive effects of songs and include different treatment conditions, outcome measures, and participant demographics, no overall conclusions can be drawn about songs' influence on writing outcomes.

4.3. Risk of bias

Risk of bias (RoB) assessment results for each included study are summarised in Table 14. The final column indicates overall weight of evidence, with 'strong' trustworthiness ratings in green, 'moderate' ratings in yellow and 'limited' in pink. The supplementary materials contain full MMAT assessment results and commentary. Key studies' methodological strengths and weaknesses are tabulated in outcome-specific categories in Section 4.2.5 above.

4.3.1. Cumulative confidence across studies

Of the 60 included studies, three received 'strong', 14 'moderate', and 43 'limited' global weight of evidence ratings. As Fig. 10 illustrates, studies with high RoB make up two thirds of included papers. Problems arose primarily in defining research questions clearly and reporting how data addressed them in adequate detail (the two screening questions); using appropriate measurements, such as standardised or validated instruments; and accounting for confounders in the design and data analysis. The largest source of bias was failing to account for confounders in two thirds of the studies (n = 43). This could be addressed by including baseline measures or allocating participants randomly at the individual level to experimental and control conditions, an allocation strategy which was not reported clearly in any included studies (see Table 6).

Most studies report songs' positive effects on their measures, but the cumulative weight of evidence is limited, as Fig. 11 illustrates. Positive effects are noted on vocabulary (receptive and productive), grammar, and speaking measures from studies with low RoB ratings, but also some neutral effects for grammar and productive vocabulary. Therefore, no overall conclusions can be drawn about the substantive linguistic effects of using songs to teach second or foreign languages to young learners in compulsory formal education. It is clear from these results that, in any future research, our confidence in understanding the effects of using songs for SLA with YLLs stands to be improved if careful methodical steps are taken to minimise the biases that have the potential to mislead us.

Table 12
Studies reporting reading measures.

	Study ID	Reading Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRESCHOOL	3. Allen-Tamai (2000)	Rhyme awareness tested by children raising a pink flag if a word rhymes, or green if it does not when told a word and asked which of the two words read aloud (supported with visuals) shares the same end sound (hold up pink or green flag for each). 30 questions with rhyming words taken from two taught nursery rhymes with implicit (nursery rhyme) or explicit (rhyming word game) conditions. Tests video recorded and researcher noted children's responses afterwards.	Claims no significant differences in mean scores between groups: children improved their rhyme awareness regardless of type of instruction. Children acquired rhyme knowledge equally well from explicit (rhyming games) or implicit (nursery rhyme) conditions, thus author claims nursery rhymes are useful as semantic material for developing L2 rhyme awareness.	Researcher-designed, non-standardised tests and testing procedure (children held up coloured flags to indicate their responses). No baseline imbalances are reported and confounding factors inadequately accounted for.
	9. Augustine (2015)	Print knowledge, definitional vocabulary, phonological awareness tested with TOPEL (Test of Preschool Early Literacy).	Claims positive effect of music treatment on overall reading scores: significant differences on print knowledge and definitional vocabulary, but not phonological awareness.	Baseline measures not reportedly fully (only p value); English L2 skills not reported clearly; no description of what happened in the intervention or whether it continued as intended.
	30. Herrera, Lorenzo, Defior, Fernandez-Smith & Costa-Giomi (2011)	Phonetic awareness, verbal memory, naming speed, name and sound letters knowledge.	Rhyme oddity task: both treatment groups outperformed the control group, with musical treatment significantly outperforming non-musical phonological training ($p < .05$) regardless of L1 status. Syllabic tapping and initial phoneme oddity task: both treatment groups outperformed controls at post-test, but it does not report if the treatment groups' mean scores were significantly different from each other. Naming task: treatment groups outperformed controls. Tamazight (L2 Spanish) learners in the music group significantly outperformed Tamazight learners in the control group.	RQs not stated (just aims/hypotheses) hence low MMAT score. Otherwise, tests were standardised; the three groups were not significantly different in terms of vocabulary, intelligence, prereading knowledge, or memory scores at the beginning of the project, and stratified random allocation to groups was used.
PRIMARY	24. Dominguez (1991)	Basic reading skills (e.g., word recognition, digraphs, end sounds, letter sounds, referents, drawing conclusions, predicting outcomes, etc.)	Only the word recognition test (1/15 tests) had a significant difference in mean scores between the treatment and control groups.	Researcher-designed instruments, non-standardised, and change modality from intervention > posttest; no cognitive ability baseline, which could be a confounding factor.
	25. Fonseca-Mora, Jara-Jiménez & Gómez-Domínguez (2015)	Early grade reading assessment (EGRA): letter name knowledge, oral reading fluency, initial sound identification.	Claim that performance of the phonological training and phonological training with music groups increased significantly compared to control group for correct letter names test, but not for correct words read in a dialogue or initial sound identification tests.	RQs not stated (just aims/hypotheses); baseline measures of musical aptitude, intelligence, reading habits, phonological awareness, parent education level, L1, but Spanish gypsy families formed the non-musical group, which is only raised in the discussion and may be a confound.
	37. LeBrun (2019)	¡Así se dice! End-of-unit test: read two paragraphs about the weather unit – 10 points for reading comprehension questions.	No significant between-groups differences in mean reading scores.	Tests taken from the textbook, so not standardised measures; ANCOVA used to control for vocab, grammar, listening and reading baselines, but other baseline measures not reported and groups were intact classes.
	57. Toscano-Fuentes & de Vega (2018)	Silent reading fluency test: spend one minute reading the text (in English), identifying and segmenting as many words as possible with a pencil.	Claim positive effect of music videos with subtitles on performance in silent reading fluency in English.	RQ not stated (just aim to use songs to improve L2 reading fluency); no baseline measures other than pretest and confounders unaccounted for; no non-song comparison group.
SECONDARY	40. Lowe (1995)	Reading comprehension – the comprehension section of the test consisted of a short text to read, after which students were asked to answer five items as 'true' or 'false' and five items which required them to fill in a blank.	Claims a significant effect of music treatment on reading comprehension: the experimental group made more progress than control group from pre- to post-test, when maths and French prior achievement are covariates.	Researcher-devised non-standardised measures of music, language, and maths achievement; no random allocation thus confounds inadequately accounted for.
			Green = positive, Yellow = mixed, Pink = negative	Green = strong, Yellow = moderate, Pink = limited

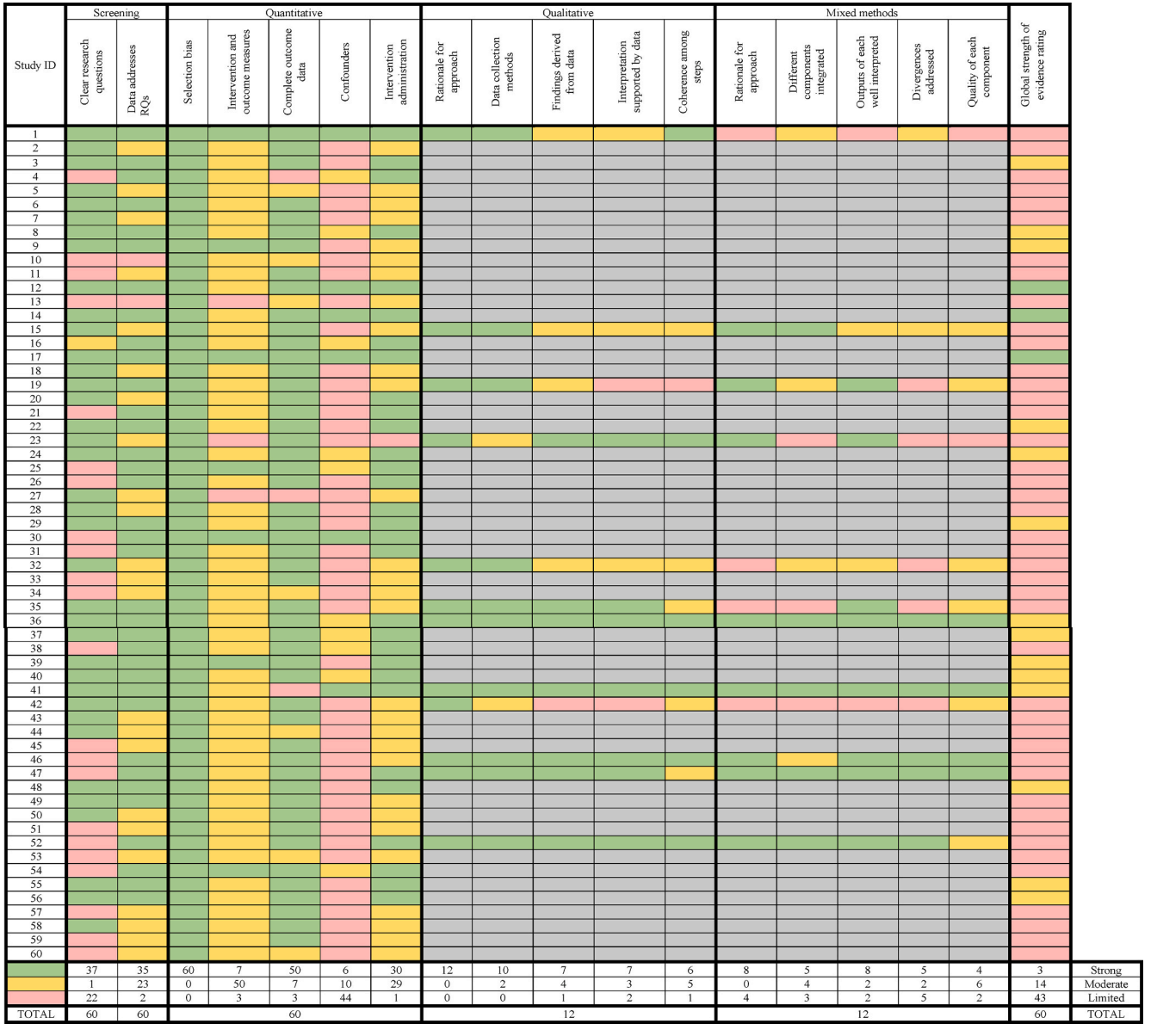
5. Discussion

Whilst support for songs' effectiveness as tools for teaching YLLs appears in peer-reviewed journals (Degrave, 2019; Paquette & Rieg, 2008; Şevik, 2011) and non-peer reviewed publications (Davanellos, 1999; Linse, 2006; Saricoban & Metin, 2000), there seems to be limited reliable evidence in either context to justify claims that using songs is especially facilitative of language learning. Critical reviews previously found few studies investigating song use with YLLs, reporting a mismatch between teacher practice and strong theoretical or empirical foundations underpinning practice (Davis, 2017; Engh, 2013; Sposet, 2008). This review's results demonstrate that research investigating songs' influence on a variety of linguistic outcomes has been accumulating since the 1970s, in diverse

Table 13
Studies reporting writing measures.

	Study ID	Writing Measure	Claim made by authors about findings Green = positive, Yellow = mixed, Pink = negative	MMAT commentary & trustworthiness indicator Green = strong, Yellow = moderate, Pink = limited
PRIMARY	11. Boey (1978)	10 marks for sentence dictation as part of end-of-year assessment.	No significant difference between experimental and control groups in their English dictation.	RQs not reported, just an aim. Unclear whether data addresses question since results are hard to follow – who is pilot, who is follow-up? Two-year study and confounders are not accounted for. Very little description of the intervention itself, so difficult to ascertain what happened.
SECONDARY	36. Klohs (1994)	Write one paragraph (scored out of 15) about an event from the weekend or from childhood. Include negatives, questions, and sentences about other people. Marked according to Semke's Communicative Rating Scale of 1 (unintelligible) to 5 (Mostly intelligible). Three French NS hired to rate the essay task.	Only predictor of success in the essay task was the previous quarter grade, not treatment condition, according to the stepwise regression model used.	Researcher-devised grammar and essay tasks; participants from three classes were matched before being randomly assigned to treatment groups but no other baselines are reported. Good integration of mixed methods data to draw conclusions.

Table 14
Risk of bias of individual studies.



Study ID	Screening		Quantitative					Qualitative					Mixed methods					Global strength of evidence rating
	Clear research questions	Data addresses RQs	Selection bias	Intervention and outcome measures	Complete outcome data	Confounders	Intervention administration	Rationale for approach	Data collection methods	Findings derived from data	Interpretation supported by data	Coherence among steps	Rationale for approach	Different components integrated	Outputs of each well interpreted	Divergences addressed	Quality of each component	
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		
25																		
26																		
27																		
28																		
29																		
30																		
31																		
32																		
33																		
34																		
35																		
36																		
37																		
38																		
39																		
40																		
41																		
42																		
43																		
44																		
45																		
46																		
47																		
48																		
49																		
50																		
51																		
52																		
53																		
54																		
55																		
56																		
57																		
58																		
59																		
60																		
	37	35	60	7	50	6	30	12	10	7	7	6	8	5	8	5	4	3	Strong
	1	23	0	50	7	10	29	0	2	4	3	5	0	4	2	2	6	14	Moderate
	22	2	0	3	3	44	1	0	0	1	2	1	4	3	2	5	2	43	Limited
TOTAL	60	60	60					12					12					60	TOTAL

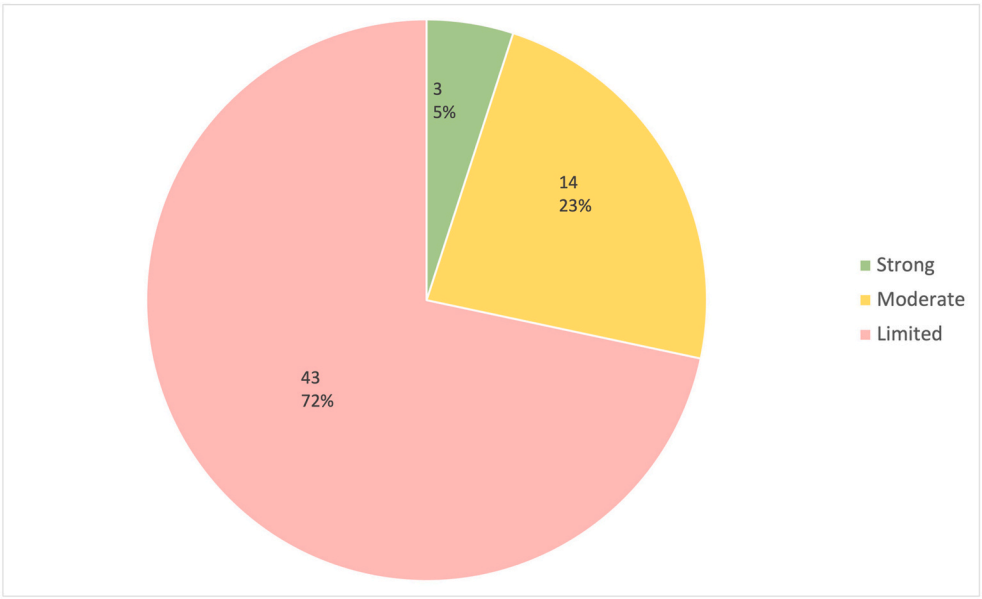


Fig. 10. Global weight of evidence ratings (MMAT).

geographical contexts, with learners from the full formal education age range of 2–18 years. The lack of evidence on the causal relationships between singing songs and substantive foreign language learning outcomes with YLLs may not be because research has not taken place (although 60 eligible studies over four decades might appear somewhat limited), but rather, because of the methodological appropriateness of that research.

Research to date often fails to reliably capture any effects of songs or measure the influence of songs on YLLs' linguistic outcomes. Research questions often do not move the field forward and are not motivated by strong theories. Research designs are often limited and opaquely reported, making it impossible to build on existing designs. In many cases, methods are not rigorous enough to deliver

Fig. 11. Reported effect of singing and weight of evidence.

the highest quality evidence and not reported transparently enough. Data are too often analysed with statistical methods relying on inference despite small sample sizes not producing generalisable or reliable inferences. This is particularly relevant where intact classes are allocated to conditions, where the class constitutes the case, not the individual. Twenty-five studies had only two intact classes (cases), which for inferential statistical purposes is limited, even if those classes are randomly allocated to conditions. Drawing causal conclusions from such studies is therefore problematic (Chalmers & Murphy, 2022). We contend that to confidently understand the substantive effects on linguistic outcomes of singing songs with YLLs, well-powered, robustly designed fair tests (e.g., randomised trials) of these approaches are needed. Resources permitting, this should be the aim for future research if we are to reliably make claims about songs' effects on FL learning outcomes.

In the absence of such research, quasi-experimental designs certainly signpost important findings for teachers and researchers, but these studies should be viewed in the context of their own methodological limitations, rather than cited as widely applicable evidence of an effect. Multiple quasi-experimental studies producing similar patterns of findings may indicate potentially fruitful avenues for future larger-scale research. However, this review found scant evidence of studies laying reliable groundwork for future research. Despite low cumulative confidence across included studies, interpretations of findings are often positively biased and lack transparent acknowledgement of their limitations.

There are several possible reasons for intervention research in this field systematically failing to achieve the highest quality-threshold. Since songs are already popular resources with language teachers (Garton et al., 2011; Harris and O'Leary, 2009) who often rely on shared experiential rather than new empirical evidence for making informed practical decisions (Borg, 2009; Bruner, 1996; Paran, 2017), it could be the case that a less critical lens is being adopted when probing the evidence base for using songs as SLA pedagogy because songs feel 'natural' to use and are intuitively appealing due to culture-based assumptions.

In turn, teachers may not be aware of needing this research because songs are part of the fabric of teaching (Hamilton & Murphy, 2023). Teachers who follow their curiosity about the evidence and investigate this valued practice are already interested in using songs or convinced of songs' practical value as language-learning tools, since they have seen how beneficial they appear to be in class: the prevailing feeling is that songs 'work' for multiple pedagogical and classroom purposes (Forster, 2006; Hamilton & Murphy, 2023; Paquette & Rieg, 2008). There is thus a positive bias in the field whereby researchers attempt to verify a prior assumption that songs 'work'.

This review found that the most trustworthy studies (those with low RoB) were equally likely to find positive or equivocal effects of singing on their outcome measures. There was, however, a considerable positive skew in claims being made by papers with high RoB. Well-conducted intervention research, which is in the minority, has yet to build up a clear picture about songs' contribution to SLA. Meanwhile, evidence reported in more numerous but less trustworthy studies continues to circulate and appears to support 'folk theory' (Bruner, 1996) about songs' efficacy.

Furthermore, included studies' theoretical frameworks and contribution to theory is generally limited. Many frame their motivation for research into using songs with YLLs in terms of songs' ubiquity in education and draw upon untested (Mitchell, Myles & Marsden, 2019) linguistic hypotheses such as Krashen's (1985) comprehensible input and affective filter, mnemonic hypotheses such as 'Song Stuck in My Head' (Murphey, 1990) or 'din' (Krashen, 1983), and research fields such as learning styles or multiple intelligences for which a unified theoretical basis and empirical substantiation are limited (Coffield et al., 2004; Waterhouse, 2006). Where both the theoretical foundations and the methodological rigour of many included studies are insubstantial, this presents a considerable challenge for the field's coherence and progression.

Additionally, few included studies build upon previous findings. For example, Davis and Fan (2016)[22] seek to resolve methodological flaws in Chou (2014)[19], Coyle and Gómez Gracia (2014), and Medina (1991)[48] by isolating songs as a variable and using adequate controls. However, papers citing Davis and Fan (2016) include pedagogical recommendations for using chants (e.g., Cedeño & Santos, 2021) but none of the included subsequent studies[12,42,56] measuring vocabulary acquisition build upon Davis and Fan's methodology or findings. Such examples of overlooking existing research evidence indicate missed opportunities to push the field forward. This may partially explain why no overall causal conclusions can be drawn from the 35 studies measuring vocabulary acquisition: too many papers begin with the question of whether songs influence vocabulary acquisition rather than building upon prior knowledge, finessing research questions and methodologies, and replicating findings in new contexts. Vocabulary knowledge is an important predictor of L2 success (Murphy, 2014) and the research could have a real impact on learning outcomes. Future studies could build on Davis and Fan (2016) by using linear mixed effects models for the data analysis that would account for data clustering (items nested in individuals) and by using an ecologically valid control (i.e., items presented in taught alternative conditions).

Promisingly, Campfield and Murphy (2013)[14] indicate a future direction for songs research by demonstrating that prosodically salient nursery rhyme input positively impacts Polish EFL learners' ability to judge English word order. A possible follow-up study could attempt to replicate these findings, adding a sung condition as well as nursery rhymes (which in their study present rhythmically salient input without melody), teasing apart the effects of prosody and melody in L2 acquisition to ascertain whether prosody's influence on learning can be enhanced by melody or whether there is no additional benefit, as found for vocabulary acquisition in Davis and Fan (2016). Further theoretical support for such an endeavour comes from lab-based word-order acquisition studies with adult L2 learners (Saksida et al., 2021) and evidence that teaching L2 prosody and suprasegmental features explicitly improves fluency and comprehensibility of L2 learners' speech (Gordon & Darcy, 2016).

Certainly, the time has come to build solid theoretical foundations for using songs in L2 contexts that are substantiated by rigorous empirical evidence. Few studies build upon prior knowledge, gathering evidence in carefully controlled conditions to answer increasingly nuanced, theoretically driven questions. Despite the number of experimental studies reviewed here, the field seems to have lost momentum, perhaps reflecting an acceptance of the 'folk theory' (Bruner, 1996) that songs 'work'. Hopefully this review provides a clear map of existing research and the state of the knowledge within this substantive area, permitting future studies to move

the field forward rather than going over the same ground.

Given songs' popularity with teachers, collaborating with practitioners to create intervention research that empirically tests intuition-driven practice and observations drawn from exploratory studies might be useful since it is important to conduct research that aligns with and underpins current practice. Teachers' long-standing cultural beliefs about songs' effectiveness need to be addressed if future research is to catalyse any change in pedagogical approaches. Research needs to be clearly signposted as exploratory or confirmatory, and reported transparently with careful attention to potential biases if practitioners are to view empirical evidence as trustworthy. The current state of the field does little to garner practitioners' confidence in research findings, whether positive or negative, and will arguably have little impact on practice, which will continue to follow its own experiential-based intuition (Bruner, 1996; Paran, 2017).

5.1. Limitations

Whilst this review sought any intervention design and had liberal inclusion criteria to gather maximum available evidence because previous reviews had found few includable papers (e.g., Davis, 2017), it lacked a quality control exclusion criterion, which may be a limitation given the high RoB ratings of many included papers. However, our intention was to ascertain the extent and nature of intervention research into using songs with YLLs, not to focus on niche effects in this field. Findings reflect broad interest in the topic despite the limited overall quality of intervention research. Including grey literature, which comprised about a third of included studies, broadens the search but is a potential limitation since theses are not peer reviewed. Considering the high RoB even of peer-reviewed papers, including grey literature does not appear to have skewed findings.

Searches only targeted three languages other than English, perhaps overlooking research from further languages. Keywords aimed for comprehensiveness, yet relevant terms may have been omitted (e.g., additional specific linguistic outcomes). The review may be biased by not acquiring five full texts, which might have provided further reliably gathered and reported findings. However, the overwhelming conclusion that intervention research into this important area of teaching practice with YLLs lacks reliability and strength would likely remain unchanged.

6. Conclusion

This systematic review investigated the extent and nature of intervention research evaluating the substantive linguistic effects of using songs to teach second or foreign languages to young learners aged 2–18 years in formal education contexts. It is worth noting here again that the experience of young language learners needs to be considered through qualitative and quantitative approaches to research. In focusing on intervention research, this review seeks to provide a comprehensive analysis of just one thread in the rich tapestry of research investigating using songs with young FL learners. The findings demonstrate interest in this field worldwide, particularly since 2009, with studies predominantly conducted in primary schools. Of the 60 included studies, over half focus on vocabulary learning as their outcome measure, followed by grammar and pronunciation.

43 studies received high RoB ratings, with systematic limitations detected in the use of unstandardised or unvalidated instruments for measuring outcomes, and lack of accountability for confounding factors, including poor baseline measures and failure to create unbiased comparison groups (or failure to compare the intervention with anything else at all). The overall weight of evidence is thus limited. Despite scant trustworthy experimental evidence in the field, many researchers make positive claims about the effectiveness of singing songs with YLLs for learning vocabulary, grammar and improving language skills. Currently, there is no clear mandate for claiming any effects (positive, neutral, or negative) on any outcome measure from the three studies with low RoB.

Since our review includes any intervention research design where linguistic outcomes were measured, due consideration of the limitations of these designs is needed. We have focused on what *can be* inferred in terms of causal links between using songs and language learning in preschool, primary and secondary educational contexts from the included studies, and found extremely limited evidence for controlled trials and reliable causal designs. That is not to say that none of the other gathered evidence has value but, given the prevalence of the causal assumptions in popular culture, we feel it is important to establish very clearly what we do and do not know. Whilst space does not permit an exhaustive account of all possible conclusions that might be drawn from included studies, this paper gives a clear mandate for further and better causal designs to be implemented, and for more reliable and transparent reporting of intervention research, and what it can (or cannot) reliably contribute to our collective knowledge in this field.

Funding declaration

The authors received no financial support for the research, authorship, and/or publication of this article. The authors report there are no competing interests to declare.

CRediT authorship contribution statement

Catherine Hamilton: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Johannes Schulz:** Writing – review & editing, Methodology, Formal analysis. **Hamish Chalmers:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Victoria A. Murphy:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Acknowledgements

With thanks to Jisoo Seo for helping to locate the Korean papers and Hyunjin Kim for assistance with their screening and data extraction, and Claire Yu Hao for helping to obtain a Chinese paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2024.103350.

Appendix A. example search strings

There were differing limits to how many search terms could be included on different databases/in different languages, as reflected in the example search strings reported in Table A1.

Table A1
Example search strings.

Language	Database	Search string
English	ProQuest Education	ab (MFL OR EAL OR ESL OR EFL OR "foreign language*" OR FL OR "second language*" OR L2 OR French OR German OR Spanish OR English OR TEFL OR TESOL) AND ab (KS1 OR KS2 OR KS3 OR KS4 OR "key stage" OR EYFS OR "early years" OR preschool OR kindergarten OR infant* OR junior* OR primary OR secondary OR elementary OR child* OR adolescent* OR "high school") AND ab ("nursery rhyme*" OR choral OR chant* OR song* OR music* OR sing*) AND ab (vocabulary OR gramma* OR phonolog* OR acquisition OR speaking OR spoken OR proficiency OR competence or skill*) NOT ab (singapore OR single* OR singular)
French	Pascal-Francis	((FLE OR anglais OR "langue étrangère" OR français OR FLS OR "langue seconde" OR allemand OR espagnol OR "langue* moderne*") AND (jeune* OR maternelle OR primaire OR collège OR élémentaire OR enfan* OR adolescent OR lycée) AND (vocabulaire OR grammaire OR phonologie OR acquisition OR compétence) AND (comptine* OR choral* OR chant OR chanson* OR chanter OR musique OR musical*)
German	Fachportal Pädagogik	(Titel: DAZ oder DAZ oder DAF oder DAF oder L2 oder SLA oder TEFL oder TESOL oder TESL oder ENGLISCH oder FRANZOESISCH oder SPANISCH oder FREMDSPRACH* oder ZWEITSPRACH* oder ZWEISPRACHIG) und (Schlagwörter: LERNER oder GRUNDSCHULE oder KIND* oder JUGENDLICH* oder GYMNASI* oder REALSCHULE oder GANZTAGSSCHULE oder GESAMTSCHULE oder HAUPTSCHULE oder FOERDERSCHULE oder SCHUELER*)) und (Freitext: LIED* oder REIM oder GESANG oder SING* oder SPRECHCHOR oder SONG oder MUSIK oder RHYTHMUS oder RHYTHMISCH oder MELODIE oder MUSIKALISCH oder MELODISCH)) und (Freitext: VOKABEL* oder GRAMMATIK oder PHONOLOGIE oder ERWERB oder LERN*)) und nicht (Freitext: SINGAPUR oder SINGLE)
Spanish	TESEO educacion.gob.es	("idioma adicional" O "lengua inglesa" O "idioma extranjero" O "lengua* extranjera*" O "secunda lengua" O "secundo idioma" O francés O "lengua castellana" O español O inglés O "lenguas modernas" O "lenguas vivas" O "idiomas modernos") Y (guardería O "jardín de infancia" O "escuela infantil" O "escuela preescolar" O "escuela secondaria" O instituto* O "escuela de primaria" O "enseñanza primaria" O "escuela elemental" O "ciclo primario" O niño* O estudiante*) Y (rimas infantiles O coral O canto* O canción* O música* O cantar) Y (vocabulario O gramática O fonologi* O adquisición O "habilidades lingüísticas" O "conocimientos lingüísticos")

Appendix B. blank data extraction form

	Item	Data	Description/Translation
General	Date form completed		dd/mm/yyyy
	ID of person extracting data		Name, email
	Reference citation		Full APA reference
	Study author contact details		Email or address
	Publication type		e.g. full report, abstract, thesis
	Document Source		Source database, website or institute
	Study funding source		
	Notes		
Study overview	Research Questions		
	Study design		e.g. RCT, observational, case study
	Study type		e.g. classroom intervention, psycholinguistic research
	Data type		Quantitative/qualitative
	Study duration		Include start date. End date, and duration if possible
	Location and language of publication		Country/language
Participants	School setting (social and educational context)		e.g. primary, secondary, public, private (if laboratory study, put n/a)
	Recruitment		How were schools/participants recruited?

(*continued on next page*)

(*continued*)

	Item	Data	Description/Translation
	Population description		Include any information regarding participants' learning disabilities, socioeconomic background, etc.
	Languages spoken		Indicate L1/L2/L3, majority/minority/foreign, and proficiency level at beginning of the study in each language, as appropriate.
	Age		What is the age range and the number at each age?
	Gender		Include gender breakdowns where available.
	Other relevant sociodemographics		
Intervention	Language of instruction		Which language was used predominantly?
	Description		What did the intervention entail?
	Duration/timing		How long did the intervention last?
	Comparison (if any)		Was any control group included in the study? If so, what distinguished them from the intervention group?
	Number of participants		n = total number of participants n = intervention group n = control group
	Class grouping		Were participants grouped in classes? Describe differences.
	How groups were generated		e.g. random allocation at individual level, cluster randomisation at class level, no report of allocation strategy, etc.
	Baseline imbalances		Any significant differences at the beginning of the study?
	Attrition		Did any participants leave the study? How many, and for what reason?
Outcomes	Outcome type		(language skills in L1, L2, etc., content knowledge, attitudes, other.)
	Outcome name		e.g. vocabulary knowledge, speaking proficiency, etc.
	Unit(s) of measure		How is outcome operationalised?
	Time points measured		How many data collections? When?
	Descriptive outcomes		Summary of outcomes and descriptive statistics
	Effect sizes		Effect sizes (if reported) or other relevant statistics

References

denotes papers included in the systematic review

* Albaladejo, S. A., Coyle, Y., & Larios, J. R. de. (2018). Songs, stories, and vocabulary acquisition in preschool learners of English as a foreign language. *System, 76*, 116–128. https://doi.org/10.1016/j.system.2018.05.002.

* Alinte, C. (2013). Teaching Grammar through Music. *The Journal of Linguistic and Intercultural Education, 6*, 7–28. https://doi.org/10.29302/jolie.2013.6.1.

* Allen-Tamai, M. (2000). *Phonological Awareness and Reading Development of Young Japanese Learners of English*. Temple University. Doctoral thesis.

* Alley, D. C. (1988). *The role of music in the teaching of listening comprehension in Spanish*. University of Georgia. Doctoral thesis.

* Al-Mosawi, F. R. A. H. (2018). Finger Family Collection YouTube Videos Nursery Rhymes Impact on Iraqi EFL Pupils' Performance in Speaking Skills. *Opción. Año, 34*, 452–474. Especial No.17(2018).

* Amiri, M., & Sobouti, F. (2016). The effect of using short stories and songs on the second language achievement of Iranian young learners. *Modern Journal of Language Teaching Methods (MJLTM), 6*(5), 401–412.

* An, G.-H.안근행 (2009). The effects of teaching English song through Korean song in vocabulary acquisition and affective attitude,우리 동요를 활용한 영어 노래 지도가 어휘력과 정의적 태도에 미치는 영향. *Journal of the Korea English Education Society,*영어교과교육, *8*(1), 37–57.

Arthur, C. (2023). Why do Songs get "Stuck in our Heads"? Towards a Theory for Explaining Earworms. *Music & Science, 6*, 205920432311645 https://doi.org/10.1177/20592043231164581.

* Au, T. K. (2013). Songs as Ambient Language Input in Phonology Acquisition. *Language Learning and Development, 9*(3), 266–277. https://doi.org/10.1080/15475441.2013.753819.

* Augustine, C. (2015). How the use of music and movement impacts the learning of reading skills by preschoolers. *Malaysian Music Journal, 4*(2), 74–90.

Bainbridge, C., Youngers, J., Bertolo, M., Atwood, S., Lopez, K., Xing, F., & Mehr, S. (2021). Infants relax in response to unfamiliar foreign lullabies. *Nature Human Behaviour, 5*, 256–264. https://doi.org/10.1038/s41562-020-00963-z

Bangerter, A., & Heath, C. (2004). The Mozart effect: Tracking the evolution of a scientific legend. *British Journal of Social Psychology, 43*, 605–623. https://doi.org/10.1348/0144666042565353

Barber, E. (1980). Language Acquisition and Applied Linguistics. *ADFL Bulletin, 12*(1), 26–32.

Becerra Vera, B., & Luna, R. M. (2013). Teaching English through music: a proposal of multimodal learning activities for primary school children. *Encuentro, 22*, 16–28.

Bedford, D. A. (1985). Spontaneous playback of the second language: A descriptive study. *Foreign Language Annals, 18*(4), 279–287. https://doi.org/10.1111/j.1944-9720.1985.tb01805.x.

* Boey, L. K. (1978). The Unified Language Project. *RELC Journal, 9*(1), 19–27.

Boland, A., Cherry, M. G., & Dickson, R. (2017). *Doing a systematic review: A student's guide* (2nd edition.). London: SAGE.

Borg, S. (2009). English Language Teachers' Conceptions of Research. *Applied Linguistics, 30*(3), 358–388. https://doi.org/10.1093/applin/amp007

Bruner, J. S. (1996). *The Culture of Education*. Cambridge, MA: Harvard University Press.

* Busse, V., Hennies, C., Kreutz, G., & Roden, I. (2021). Learning grammar through singing? An intervention with EFL primary school learners. *Learning and Instruction, 71*, Article 101372. https://doi.org/10.1016/j.learninstruc.2020.101372.

* Caleya, M. F., Nieto, M., & Espejo, A. (2013). Music, poetry and fun activities in English teaching: an early childhood education experience. *EDULEARN13: 5th International Conference on Education and New Learning Technologies, 0*(0), 1473–1481.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Company.

* Campfield, D. E., & Murphy, V. A. (2013). The influence of prosodic input in the second language classroom: does it stimulate child acquisition of word order and function words?. *The Language Learning Journal, 45*(1), 81–99. https://doi.org/10.1080/09571736.2013.807864.

Campfield, D. E., & Murphy, V. A. (2014). Elicited imitation in search of the influence of linguistic rhythm on child L2 acquiition. *System, 42*, 207–219. http://10.1016/j.system.2013.12.002.

Cedeño, C., & Santos, L. (2021). Chants in EFL Vocabulary Instruction with Young Learners: Potential, Composition and Application. *JELTL (Journal of English Language Teaching and Linguistics), 6*(1), 153–165.
* Chae, Y., & Yoon, E. (2013). The Effects of the Songs of Children's Literature on the Primary School Students' Long-term Memory, Grammar Learning, and Affective Domains,영어동화노래수업이 장 · 단기 기억과 문법습득 및 정의적 영역에 미치는 효과. *Primary English Education, 초등영어교육, 19*(2), 241–270.
Chalmers, H. (2016). Can Education Learn from Evidence-Based Medicine? *Centre for Evidence Based Medicine*. Retrieved February 22, 2023. From https://ebmlive.org/can-education-learn-from-evidence-based-medicine/.
Chalmers, H., Brown, J., & Koryakina, A. (2023). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2022-0190
Chalmers, H., & Murphy, V. A. (2022). Multilingual learners, linguistic pluralism and implications for education and research. In E. Macaro, & R. Woore (Eds.), *Debates in Second Language Education*. New York: Routledge. https://doi.org/10.4324/9781003008361-6.
* Cheippe, E. (2012). *La voie musicale pour remédier aux difficultés de prononciation des voyelles de l'allemand dans des textes lus: expérimentation dans une classe bilingue: analyse acoustique*. Université de Strasbourg. Doctoral thesis.
* Chen, J.-J. (2011). *The effects of music activities on English pronunciation and vocabulary retention of fourth-grade ESOL (English for Speakers of Other Languages) students in Taiwan*. University of Florida. Doctoral thesis.
* Chiang, M. (2003). *The effect of chanting activities on the comprehension of English of first graders and college freshmen in Taipei*. Texas A&M University-Kingsville. Doctoral thesis.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.
* Chou, M. (2014). Assessing English vocabulary and enhancing young English as a Foreign Language (EFL) learners' motivation through games, songs, and stories. *Education 3–13, 42*(3), 284–297. https://doi.org/10.1080/03004279.2012.680899.
Csizér, K., Albert, Á., & Piniel, K. (2022). Editorial: Introduction to the special issue on conducting research syntheses on individual differences in SLA. *Studies in Second Language Learning and Teaching, 12*(2), 157–171.
Cochrane Effective Practice and Organisation of Care (EPOC). (2017). Data collection form. Retrieved from https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/good_practice_data_extraction_form.doc. (Accessed 6 December 2021).
Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning. A systematic and critical review*. London: Learning Skills and Research Centre, Department for Education.
Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. Los Angeles: SAGE.
Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Chicago: Rand McNally College.
Coyle, Y., & Gómez Gracia, R. (2014). Using Songs to Enhance L2 Vocabulary Acquisition in Preschool Children. *ELT Journal, 68*(3), 276–285. https://doi.org/10.1093/elt/ccu015
Crosswhite, J. (1996). *Effect of music instruction on language development of preschool children*. University of North Carolina. Doctoral thesis.
* Cruz-Cruz, M. L. (2005). *The effects of selected music and songs on teaching grammar and vocabulary to second grade English language learners*. Kingsville: Texas A&M University. Doctoral thesis.
Davanellos, A. (1999). Songs. *English Teaching Professional, 13*, 13–15.
Davis, G. M. (2017). Songs in the Young Learner Classroom: A Critical Review of Evidence. *ELT Journal, 71*(4), 445–455. https://doi.org/10.1093/elt/ccw097
* Davis, G. M., & Fan, W. (2016). English Vocabulary Acquisition Through Songs in Chinese Kindergarten Students. *Chinese Journal of Applied Linguistics, 39*(1), 59–71. https://doi.org/10.1515/cjal-2016-0004.
Degrave, P. (2019). Music in the Foreign Language Classroom: How and Why? *Journal of Language Teaching and Research, 10*(3), 412–420. https://doi.org/10.17507/jltr.1003.02
* Diakou, M. (2014). Using Songs to Enhance Language Learning and Skills in the Cypriot Primary MFL Classroom. *EdD thesis*. The Open University.
* Dominguez, D. (1991). *Developing language through a musical program and its effect on the reading achievement of Spanish-speaking migrant children*. Wester Michigan University. Doctoral thesis.
Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised*. Circle Pines, MN: American Guidance Service, Inc.
Engh, D. (2013). Why use music in English language learning? A survey of the literature. *English Language Teaching, 6*(2), 113–127. https://doi.org/10.5539/elt.v6n2p113
Fonseca-Mora, M. C. (2000). Foreign language acquisition and melody singing. *ELT Journal, 54*(2), 146–152. https://doi.org/10.1093/elt/54.2.146
* Fonseca-Mora, M. C., Jara-Jiménez, P., & Gómez-Domínguez, M. (2015). Musical plus phonological input for young foreign language readers. *Frontiers in Psychology, 6*, 286. https://doi.org/10.3389/fpsyg.2015.00286.
Forster, E. (2006). The value of songs and chants for young learners. *Encuentro, 16*, 63–68.
Franco, F., Suttora, C., Spinelli, M., Kozar, I., & Fasolo, M. (2021). Singing to infants matters: Early singing interactions affect musical preferences and facilitate vocabulary building. *Journal of Child Language*, 1–26. https://doi.org/10.1017/s0305000921000167
Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. NY, USA: Basic Books.
Garton, S., Copland, F., & Burns, A. (2011). *Investigating Global Practices in Teaching English to Young Learners, 11*. London: British Council: ELT Research Papers, 01.
Geisler, P. (2008). *Musikorientiertes Lernen im Englisch-Unterricht der Grundschule*. Pädagogische Hochschule Freiburg. Doctoral thesis.
Gervain, J., Christophe, A., & Mazuka, R. (2020). Prosodic Bootstrapping. In C. Gussenhoven, & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 563–573). Oxford, UK: Oxford University Press.
Gil, D. G., & Azcune, B. L. (2012). Flamenco and new technologies: the music classroom as a context for the integration of the gypsy group. *Publicaciones, 42*, 121–132.
Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition, 1*, 3–55.
* Good, A. J., Russo, F. A., & Sullivan, J. (2015). The efficacy of singing in foreign-language learning. *Psychology of Music, 43*(5), 627–640. https://doi.org/10.1177/0305735614528833.
Gorard, S. (2003). *Quantitative methods in social sciences research*. New York; London: Continuum.
Gorard, S. (2013). *Research design: creating robust approaches for the social sciences*. London; Thousand Oaks: SAGE.
Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners. *Journal of Second Language Pronunciation, 2*(1), 56–92.
* Gorjian, B., Hayati, A., & Barazandeh, E. (2012). An evaluation of the effects of art on vocabulary learning through multi-sensory modalities. *Procedia Technology, 1*, 345–350. https://doi.org/10.1016/j.protcy.2012.02.072.
Gough, D., Oliver, S., & Thomas, J. (2012). *An Introduction to Systematic Reviews*. London: SAGE.
Guerrero, M. C. M. (1987). The Din Phenomenon: Mental Rehearsal in the Second Language. *Foreign Language Annals, 20*(6), 537–548. https://doi.org/10.1111/j.1944-9720.1987.tb03053.x.
* Haghverdi, H. R. (2015). The Effect of Song and Movie on High School Students Language Achievement in Dehdasht. *Procedia – Social and Behavioral Sciences, 192*, 313–320. https://doi.org/10.1016/j.sbspro.2015.06.045.
* Hakozaki, Y., & Nakagawa, Y. (2020). Teaching stress-timed rhythm of English at the Japanese elementary school level: focusing on the effects of using chants. *Asian EFL Journal Research Articles, 27*(2), 173–201.
Hamilton, C., & Murphy, V. A. (2023). Folk pedagogy? Investigating how and why UK early years and primary teachers use songs with young learners. *Education*, 3–13. https://doi.org/10.1080/03004279.2023.2168132
Harris, J., & O'Leary, D. (2009). A third language at primary level in Ireland: an independent evaluation of the modern languages in primary schools initiative. In M. Nikolov (Ed.), *Early Learning of Modern Foreign Languages. Processes and outcomes* (pp. 1–14). Bristol, Buffalo, Toronto: Multilingual Matters.
* Herrera, L., Lorenzo, O., Defior, S., Fernandez-Smith, G., & Costa-Giomi, E. (2011). Effects of phonological and musical training on the reading readiness of native- and foreign-Spanish-speaking children. *Psychology of Music, 39*(1), 68–81. https://doi.org/10.1177/0305735610361995.
Higgins, J. P. T., Altman, D. G., Gøtzsche, P., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ, 343*, d5928. https://doi.org/10.1136/bmj.d5928

Hong, Q. N., & Pluye, P. (2019). A Conceptual Framework for Critical Appraisal in Systematic Mixed Studies Reviews. *Journal of Mixed Methods Research, 13*(4), 446–460. https://doi.org/10.1177/1558689818770058

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018*. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. Retrieved from http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf. (Accessed 2 May 2022).

* Hsu, H. (2009). *The effect of rhythmic teaching methods for kindergarten EFL students in Taiwan*. University of Mississippi. Doctoral thesis.

Isaacs, T., & Chalmers, H. (2023). Reducing 'avoidable research waste' in applied linguistics research: lessons from healthcare research. *Language Teaching*, 1–18. https://doi.org/10.1017/S0261444823000411

* Jarvis, S. (2013). How effective is it to teach a foreign language in the Foundation Stage through songs and rhymes?. *Education 3-13, 41*(1), 47–54. https://doi.org/10.1080/03004279.2012.710099.

* Jeong, Y.-J., & Kim, J.-O.김정옥 (2014). A Study of English-Teaching Model through Stories and Songs,영어동화와 노래를 결합한 정의적 영어수업모형의 적용 효과. *Wonkwang Journal of Humanities,*열린정신 인문학 연구*"J, 15*(2), 57–75.

Joyce, M. F. (2011). *Vocabulary acquisition with kindergarten children using song picture books*. Massachusetts, USA: Northeastern University. Doctoral thesis.

Kaminski, A. (2016). The Use of Singing, Storytelling and Chanting in the Primary EFL Classroom: Aesthetic Experience and Participation in FL Learning. *Doctoral thesis, University of Swansea*. https://doi.org/10.23889/suthesis.54359

* Kim, J.-S., & Kang, M.-K. (2015). The Effects of Improving English Listening Skills of High School Students with a Lower Level through Pop Song Hunminglish Pronunciation (PSHP) Practice. *Advanced Science and Technology Letters, 92*(Education 2015), 41–45.

* Kim, Y., & Park, J.-E.김양희 (2012). Analysis on English vocabulary acquisition by accomplishment levels with an integrated teaching model for English and music through songs,노래를 활용한 영어 · 음악 통합 수업에서 성취 수준별 영어 어휘 습득 분석. *Primary English Education, "*초등영어교육*, 18*(3), 31–63.

* Klohs, L. M. (1994). *Use of mnemonic strategies to facilitate written production of a second language by high school French students*. University of Minnesota. Doctoral thesis.

Krashen, S. (1983). The Din in the Head, Input, and the Language Acquisition Device. *Foreign Language Annals, 16*(1), 41–44.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Harlow: Longman.

* LeBrun, C. (2019). *The Effects of Music-Infused Instruction on Student Achievement in Secondary school Spanish*. University of South Dakota. Doctoral thesis.

* Legg, R. (2009). Using Music to Accelerate Language Learning: An Experimental Study. *Research in Education, 82*(1), 1–12. https://doi.org/10.7227/rie.82.1.

Lehman, L. (2019). *Oats, peas and beans, and early literacy skills grow: A program evaluation of education through music*. New York: Alfred University. Doctoral thesis.

* Leśniewska, J., & Pichette, F. (2016). Songs vs. Stories: Impact of Input Sources on ESL Vocabulary Acquisition by Preliterate Children. *International Journal of Bilingual Education and Bilingualism, 19*(1), 18–34. https://doi.org/10.1080/13670050.2014.960360.

Linse, C. (2006). Using favorite songs and poems with young learners. *English Teaching Forum, 44*(2), 38–42.

Lonie, D. (2010). *Early Years Evidence Review: Assessing the Outcomes of Early Years Music Making*. London: Youth Music. Retrieved from Youth Music website: https://network.youthmusic.org.uk/sites/default/files/uploads/research/Early_years_evidence_review_2010.pdf. (Accessed 5 August 2021).

* Lowe, A. S. (1995). *The effect of the incorporation of music learning into the second-language classroom on the mutual reinforcement of music and language*. University of Illinois at Urbana-Champaign. Doctoral thesis.

* Ludke, K. (2010). *Songs and singing in foreign language learning*. University of Edinburgh. Doctoral thesis.

* Luo, S. (2019). Influence of Singing English Songs on Vocabulary Learning by Senior School Students in Guangzhou. *International Journal of Information and Education Technology, 9*(11), 843–848.

* Ma, S. (2004). English Education Activities and English Story Recall Using Story Songs., 이야기노래 (story songs) 를 활용한 영어교육활동과 유아의 영어이야기회상. *Early Childhood Education Research & Review,* 유아교육학논집*, 8*(2), 57–75.

* Madani, D., & Nasrabadi, M. M. (2016). The effect of songs on vocabulary retention of preschool young English language learners. *International Journal of Research Studies in Language Learning, 6*(3), 63–72. https://doi.org/10.5861/ijrsll.2016.1562.

* Mamdouh, M. (2017). La canción francófona, una herramienta eficaz en el proceso de enseñanza-aprendizaje de la lengua francesa. Thélème. *Revista Complutense de Estudios Franceses, 32*(2), 221–238. https://doi.org/10.5209/thel.54572.

* McCormack, B. A., & Klopper, C. (2016). The potential of music in promoting oracy in students with English as an additional language. *International Journal of Music Education, 34*(4), 416–432.

* McCormack, B. A., Klopper, C., Kitson, L., & Westerveld, M. (2018). The potential for music to develop pronunciation in students with English as an Additional Language or Dialect (EAL/D). *Australian Journal of Music Education, 52*(1), 43–50.

* Medina, S. L. (1991). *The effect of a musical medium on the vocabulary acquisition of limited English speakers*. University of Southern California. Doctoral thesis.

Mitchell, R., Myles, F., & Marsden, E. (2019). *Second Language Learning Theories*. London: Routledge.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

* Moradi, F., & Shahrokhi, M. (2014). The effect of listening to music on Iranian children's segmental and suprasegmental pronunciation. *English Language Teaching, 7*(6), 128–142.

Morgan, J. L., & Demuth, K. (1996). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. New Jersey, USA: Lawrence Erlbaum Associates.

Murphey, T. (1990). The Song Stuck in My Head Phenomenon: A Melodic Din in the LAD? *System, 18*(1), 53–64. https://doi.org/10.1016/0346-251X(90)90028-4

Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford, UK: Oxford University Press.

Murphy, V. A., & Castillo, J. (2013). Modality, Vocabulary Size and Question Type as Mediators of Listening Comprehension Skill. *Contemporary Foreign Languages Studies, 396*(12), 15–30.

* Muzammil, L., & Andy, A. (2019). Can Young Learners Utilize Cartoon Picture and Song To Learn? A teaching model. *Proceedings of the 3rd Asian Education Symposium (AES 2018)* (pp. 512–517). https://doi.org/10.2991/aes-18.2019.115.

* Navarro, K. S., Quiroga, C., & Diaz, C. (2018). English pronunciation for first year primary school students: a didactic sequence implementation for its improvement. *Revista Comunicación, Año, 39*(27), 108–121, 1.

Newcomer, X., & Hammill, X. (1997). *Test of Language Development-Primary* (3rd Edition). Texas, USA: Pro-Ed. TOLD:P-3.

Nunan, D., Heneghan, C., & Spencer, E. A. (2018). Catalogue of bias: allocation bias. *BMJ Evidence-Based Medicine, 23*(1), 20–21. https://doi.org/10.1136/ebmed-2017-110882

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan — a web and mobile app for systematic reviews. *Systematic Reviews, 5*(210). https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(71).

Paquette, K. R., & Rieg, S. A. (2008). Using music to support the literacy development of young English language learners. *Early Childhood Education Journal, 36*(3), 227–232. https://doi.org/10.1007/s10643-008-0277-9

Paran, A. (2017). 'Only connect': researchers and teachers in dialogue. *ELT Journal, 71*(4), 499–508. https://doi.org/10.1093/elt/ccx033

Parr, P. C., & Krashen, S. D. (1986). Involuntary rehearsal of second language in beginning and advanced performers. *System, 14*(3), 275–278. https://doi.org/10.1016/0346-251x(86)90022-9

Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences*. Oxford, UK: Blackwells.

Pluye, P., Gagnon, M. P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies, 46*(4), 529–546.

* Priester, M. (2011). *Using Song Lyrics in the Preschool ESL Classroom to Assist Students' English Vocabulary Retention and Use*. Caldwell College. Master's Thesis.

Rauscher, F. H., Shaw, G. L., & Ky, C. N. (1993). Music and spatial task performance. *Nature, 365*(6447), 611, 611 https://doi.org/10.1038/365611a0.
Richter, K. W. (2021). *Educational outcomes in multilingual CLIL school settings: A systematic review. Master's dissertation*. University of Oxford. Retrieved from https://ora.ox.ac.uk/objects/uuid:52221dda-7655-4771-ad7b-66f8c3ee23ca. (Accessed 6 January 2022).
Román-Caballero, R., Vadillo, M. A., Trainor, L. J., & Lupiáñez, J. (2022). Please don't stop the music: A meta-analysis of the cognitive and academic benefits of instrumental musical training in childhood and adolescence. *Educational Research Review, 35*, Article 100436. https://doi.org/10.1016/j.edurev.2022.100436
Saksida, A., Flo, A., Guedes, B., Nespor, M., & Garay, M. P. (2021). Prosody facilitates learning the word order in a new language. *Cognition, 213*, Article 104686. https://doi.org/10.1016/j.cognition.2021.104686
Sala, G., & Gobet, F. (2020). Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Memory & Cognition, 48*(8), 1429–1441. https://doi.org/10.3758/s13421-020-01060-2.
* Santos Jimenez, O. C., Gallegos Ruiz, A., & Gomez Hermosa, C. (2017). Application of children's songs for the learning of the French language vocabulary in children of the initial level of the San Antonio de Padua Educational Institution, Chosica, Lima, Perú, 2016. *Revista Inclusiones, 4*(4), 189–204.
Saricoban, A., & Metin, E. (2000). Songs, verse and games for teaching grammar. *The Internet TESOL Journal, 6*(10). Retrieved from http://iteslj.org/Techniques/Saricoban-Songs.html. (Accessed 7 May 2021).
Schoepp, K. (2001). Reasons for using songs in the ESL/EFL classroom. *The Internet TESOL Journal, 7*(2). Retrieved from http://iteslj.org/Articles/Schoepp-Songs.html. (Accessed 8 May 2021).
Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. A. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education, 11*(2). https://doi.org/10.1002/rev3.3413
* Schunk, H. A. (1999). The Effect of Singing Paired with Signing on Receptive Vocabulary Skills of Elementary ESL Students. *Journal of Music Therapy, 36*(2), 110–124. https://doi.org/10.1093/jmt/36.2.110.
Şevik, M. (2011). Teacher views about using songs in teaching English to young learners. *Educational Research and Review, 6*(21), 1027–1035.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalised casual inference*. Boston: Houghton-Mifflin.
Shearer, C. B. (2020). A resting state functional connectivity analysis of human intelligence: Broad theoretical and practical implications for multiple intelligences theory. *Psychology & Neuroscience, 13*(2), 127–148. https://doi.org/10.1037/pne0000200.
* Siebring, M. F. (2004). *The effectiveness of a systematic approach based on songs to prevent and correct errors in elementary core French*. Nipissing University. Master's thesis.
Slavin, R. E. (1986). Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews. *Educational Researcher, 15*(9), 5–11. https://doi.org/10.3102/0013189x015009005
Sposet, B. (2008). *The role of music in second language acquisition: a bibliographical review of seventy years of research, 1937–2007*. NY, USA: The Edwin Mellen Press.
Sterne, J. A., Hernán, M. A., Reeves, B. C. M., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919. https://doi.org/10.1136/bmj.i4919
Thain, L. A. (2010). Rhythm, music and young learners: A winning combination. In A. M. Stoke (Ed.), *JALT2009 Conference Proceedings* (pp. 407–416). Tokyo: JALT.
* Tomczak, E., & Lew, R. (2019). "The Song of Words": Teaching multi-word units with songs. *The Southeast Asian Journal of English Language Studies, 25*(4), 16–33.
* Toscano-Fuentes, C. M., & de Vega, C. J. (2018). Vídeos musicales en el aula de inglés de primaria para la mejora de la fluidez lectora / Music videos in primary English class for the improvement of reading fluency. *TEJUELO. Didáctica De La Lengua Y La Literatura. Educación, 28*, 43–66. https://doi.org/10.17398/1988-8430.28.43.
Trehub, S. E., & Trainor, L. (1998). Singing to infants: Lullabies and play songs. *Advances in Infancy Research, 12*, 43–78.
Trehub, S. E., Trainor, L. J., & Unyk, A. M. (1993). Music and speech processing in the first year of life. *Advances in Child Development and Behavior, 24*, 1–35.
Váradi, J. (2022). A Review of the Literature on the Relationship of Music Education to the Development of Socio-Emotional Learning. *Sage Open, 12*(1). https://doi.org/10.1177/21582440211068501
Walker, R. (2006). Going for a Song. *English Teaching Professional, 43*, 19–21.
* Wang, Y. (2005). A study of the effects of teaching English grammar with English songs in junior high schools. *Master's thesis*. Beijing Normal University.
Waterhouse, L. (2006). Multiple Intelligences, the Mozart Effect, and Emotional Intelligence: A Critical Review. *Educational Psychologist, 41*(4), 207–225.
Wermke, K., & Mende, W. (2009). Musical elements in human infants' cries: In the beginning is the melody. *Musicae Scientiae, 13*(2_suppl), 151–175. https://doi.org/10.1177/1029864909013002081
Wermke, K., & Mende, W. (2016). From melodious cries to articulated sounds: Melody at the root of language acquisition. In M. C. Fonseca-Mora, & M. Gant (Eds.), *Melodies, Rhythm and Cognition in Foreign Language Learning* (pp. 24–47). Newcastle: Cambridge Scholars Publishing.
Werner, V. (2020). "Song-Advantage" or "Cost of Singing"? A Research Synthesis of Classroom-based Intervention Studies Applying Lyrics-based Language Teaching (1972–2019). *Journal of Second Language Teaching and Research, 8*(1), 138–170.
Willis, S., Neil, R., Mellick, M. C., & Wasley, D. (2019). The Relationship Between Occupational Demands and Well-Being of Performing Artists: A Systematic Review. *Frontiers in Psychology, 10*, 393. https://doi.org/10.3389/fpsyg.2019.00393
* Yousefi, A. (2014). The Effect of Modern Lyrical Music on Second Language Vocabulary Acquisition. *Mediterranean Journal of Social Sciences, 5*(23), 2583–2586. https://doi.org/10.5901/mjss.2014.v5n23p2583.
* Zhaku-Kondri, B. (2014). Using Song Lyrics in the Classroom: Assessing the Utility of Song Lyrics for the Acquisition of a Foreign Language. In *Proceedings of INTCESS14 – International Conference on Education and Social Sciences Proceedings, 1* pp. 1201–1210). California Folklore Quarterly, 4.
Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (Eds.). (2020). *Systematic Reviews in Educational Research Methodology, Perspectives and Application*. Wiesbaden: Springer.