Investigating the effects of whole-class singing activities on linguistic outcomes of young foreign language learners in English primary schools.

Catherine Hamilton

M.Sc. Applied Linguistics and Second Language Acquisition, University of Oxford

M.A. English and Modern Languages, University of Oxford

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Education

St Anne's College, University of Oxford

Supervised by Professor Victoria Murphy and Dr Hamish Chalmers

March 2025

Word count: 86,429

We are all apt to be so delighted by the fact that children do get absorbed in play or fun in a language classroom that we tend to assume something worthwhile must be going on.

S. Rixon (1991)

I kan namoore expounde in this mateere,
I lerne song, I kan but smal grammere.

Chaucer, *The Prioress's Tale* (lines 83–84)

One trouble with questions whose answers are self-evident is that investigators rarely collect the evidence to see if they pan out in practice. […] What is correct about such a position is by no means obvious, and therefore deserves serious study rather than acceptance as a background fact in our field.

L. Gleitman (1990)

Acknowledgements

I've been hugely lucky throughout the process of creating this thesis to be supported by a wide cast of dear friends, colleagues and family members, all playing their roles (whether they were aware of it or not!) in this somewhat selfish production.

First and foremost, I would like to thank the two greatest supervisors on earth, Victoria Murphy and Hamish Chalmers, for your intellectual rigour, keen sense of what makes 'good' science, and for sharing your expertise so generously. Perhaps more importantly, thank you for being the all-round wonderful humans you are, and for your friendship and support during the highs and lows of the last four years. I have grown such a lot through your care. Every meeting, conference and written draft we have worked on together has brought me joy and I look forward to continuing our collaborative larks.

Huge thanks to the two (anonymous) schools and the Year 3 teachers and support staff who kindly invited me to work with them during the second phase of this research. Thanks for welcoming me and being such incredible colleagues in the planning and execution stages. There is no data without participants and what wonderful participants the children were. Eager and keen to learn French, and delightful to work with. Merci à tous for your participation.

The Department of Education in Oxford is a dream environment to work in and I am hugely grateful for the collegial and collaborative ethos there. Thank you to all three cohorts of ALSLA students for bringing structure to my weeks (I do love a timetable!), and for making me push myself to be the best teacher I can be, since that is my favourite way of learning. My own DPhil cohort is full of the kindest and loveliest fellow students: dear friends, thank you for the shared seminars, lunches, coffees, holidays, conference trips, Christmasses, walks, karaoke nights, formal halls, creative endeavours, cakes, and conversations. Thanks to my Applied Linguistics colleagues for your encouragement and facilitation of my continued teaching and learning. It is an honour to work with you all. Special thanks to Robert Woore: co-teaching has been a real pleasure, and I'm grateful for your insightful feedback on our stats module as well as on this thesis. Huge thanks to Liz Wonnacott for intellectual support through transfer and confirmation, particularly for pushing me forward in learning R and Bayes. Thanks also for your moral support: I appreciate sharing this journey with someone who 'gets' the juggle and brilliance of work/motherhood. Thanks to Sara Ratner: sometimes the universe presents you with a new person who feels like an old friend, and I'm very grateful for your kindess and care. And

thanks to dear, generous and lovely Faidra Faitaki: you taught me statistics (how can I ever thank you for helping me take those first tentative steps!) and shared your formidable research expertise, love of creativity, office, and emergency biscuits with me. I look forward to our continued arts-based adventures. Every day in the department brings conversations with incredible people, all of which have helped shaped my thinking and thus this thesis. Thank you to Kathy Sylva, Sandra Mathers, Daniella Singh, Fiona Jelley, Laura Molway and Cathy Scutt (and the whole team of librarians – utterly lost without you!), Charlotte Ryland, Eowyn Crisfield, and everyone who I have not named but enjoyed the expansive life of our department with. It is the best place I have ever had the privilege to work and learn.

Even more lucky am I that, after such fulfilling and rich days at work, I can return home to my beautiful family. Thanks to my parents, Jean and Martin, for encouraging my word-smithing from the earliest age: my life-long love of learning languages began with your support. Thanks to my brother Tom and SIL Katie for the deep conversations about primary pedagogy and music, snatched in between park trips and snack-times with our clans. Endless thanks to my husband Paul, the very best of men, for being the greatest support and source of strength through my most challenging life moments, and for being my greatest champion in the good bits. You have patiently listened while I sorted out many tangled thought processes, and often see clearly when I cannot. This whole thesis rests on your shoulders, as does my head at the end of long days of toiling over it. And finally, thank you to our wonderful children for being such brilliant linguists and getting excited about words too; for letting me bring my sentences on holiday during the data analysis phase (and for humming *The Final Countdown* on repeat when I got to the last ten!); and for helping me keep things in perspective. Danke Dylan, you always make me laugh at myself, which I *really* need sometimes. Gracias Arwen, you inspire me to face every day with a quiet, calm inner strength and grace that I admire hugely. Merci Quinn, for your patience, epic cuddles, and how you once said when I'd had a difficult day: "Isn't your job just writing sentences?" I replied that it was, but some sentences are quite stressful to write. And I'll never forget you gently explaining that sometimes you forget the full stops too, but I need not worry because "we can just put them in afterwards." Home is wherever we five are together.

Finally, thanks to past me for holding the revolving door open and letting myself back into the room. I could very easily have given up during the hard bits, but I didn't, and now I have a large book to jam the door open with. Keep going, little wordsmith. And keep enjoying the journey, one step at a time.

For Dylan, Arwen and Quinn.

And for Alex, my much-missed brother.

My daily reminder that education has the power to change the world,

but so too does love, kindness, and good humour.

Abstract

Cultural beliefs in songs' effectiveness for teaching young foreign language learners are common, despite scant credible empirical evidence to support such intuitions. This two-phase thesis investigates the empirical evidence underpinning teachers' pedagogical choices regarding the use of songs in foreign language teaching for achieving linguistic outcomes. It then proposes a coherent and systematic way forward for research evidence to better support teachers' judgements.

Phase 1 systematically reviews intervention research comparing the effects of using songs to other pedagogical tools on the linguistic outcomes of young learners in formal second or foreign language education. This phase appraises 60 studies conducted from 1978–2021, finding the cumulative evidence weak and inconclusive across all measures.

Phase 2 builds on the findings of the systematic review by implementing an intervention study with 96 beginner French learners (age 7–8) in two primary schools in England to assess the substantive effects of singing on oral language development through an elicited imitation task. It adopted a randomised controlled trial design to compare the effects on this outcome of presenting French through songs, chants or stories with each other and with an active control. Participants received 242 minutes of French input in 11 lessons over three weeks. Their elicited imitation task performance was measured at pretest, posttest and delayed posttest. Data were transcribed, scored from 0 (omission) to 5 (exact imitation), and analysed using a cumulative link mixed model. Performance improved over time for all groups, with significant interaction effects between song and control groups at posttest, and between all treatment groups and control at delayed posttest. No significant interaction effects were detected between treatment groups. Bayes factors were also calculated, finding insufficient evidence to support the null hypothesis of no difference between conditions.

This thesis empirically interrogates longstanding positive bias toward believing songs constitute effective SLA pedagogy. It contributes empirical evidence on the effects of songs

on young learners' linguistic outcomes and enhances our understanding of using songs with beginner language learners in England's primary schools, both under-researched areas. The findings have implications for teaching and future research on using songs to teach young language learners.

List of figures

List of tables

List of abbreviations

AD	adult-directed
ADS	adult-directed speech
B	Bayes factor(s)
CLM	cumulative link model
CLMM	cumulative link mixed model
DfE	Department for Education
DfES	Department for Education and Society
EFL	English as a foreign language
EU	European Union
FL	foreign language(s)
FonF	focus on form
ID	infant-directed
IDS	infant-directed speech
KS2	Key stage 2
L1	first language
L2	second language
M	mean
MFL	modern foreign language(s)
PBH	prosodic bootstrapping hypothesis
PLN	Primary Languages Network
PTH	prosodic transfer hypothesis
QCA	Qualifications and curriculum authority
RR(s)	robustness region(s)
SD	standard deviation
SE	standard error
SEN	special educational needs
SLA	second language acquisition
SSIMH	song stuck in my head
YLL(s)	young language learner(s)

Chapter 1

Introduction[1]

1.1 Background to the problem

Jolly (1975) wrote a short paper recognising the 'potential benefits' (p.11) of using songs for teaching foreign languages. Based on a handful of studies and classroom observations, the paper outlines the similarities between speech and song, and how infants' responsiveness to rhythm is an essential part of first language acquisition. Jolly proposes that language teachers might use songs for increasing students' motivation (or relieving boredom associated with drilling techniques); acquiring grammatical structures, vocabulary and idiomatic expressions; and exposing students to the target language culture. Jolly concludes that, having raised awareness of songs' *potential benefits* for foreign language (FL) learning, "our intuitive feelings will remain only ideas unless they are, in some way, proven by means of study and experimental research" (Jolly, 1975: 14). Fifty years on, the field has made little progress towards achieving that aim. While the intuitive feelings of teachers remain positive about the potential role of songs in FL learning, evidence (and especially evidence from well-conducted experimental research) that would support those feelings is notable by its scarcity.

Songs are popular resources with teachers of young language learners (YLLs)[2] worldwide (Linse, 2006; Şevik, 2011). Using songs to teach languages is often assumed by teachers to have educational and linguistic benefits, particularly with younger learners

[1] Note that parts of this chapter are adapted from Hamilton, Schulz, Chalmers and Murphy (2024). In that paper, the first author conceived and designed the study, collected the data, conducted the analysis, and wrote and edited the paper. The second author screened 10% of the inclusion decisions and contributed to the quality appraisal, to minimise bias as per good practice in systematic reviews, and edited an early draft of the paper. The third and fourth authors supported the conceptualisation of the paper, and reviewed and edited the early and final drafts.

[2] G.Ellis (2014) defines 'young learners' as children under the age of 12 years (as opposed to all pre-adulthood learners under 18 years old). This definition is important because children under the age of 12 are at the centre of the 'younger the better' debate in language learning (Murphy, 2014).

(Hamilton & Murphy, 2023). A large-scale survey of 4,696 English language teachers of YLLs from 144 countries found that 67% of respondents used songs often, or every lesson, compared to 42% reading stories as frequently (Garton, Copland, & Burns, 2011). A survey conducted among 270 schools in Ireland (Harris & 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). Singing songs has been associated with supporting social and emotional development (for a review see Váradi, 2022) and songs have been observed for their capacity to engage and motivate YLLs (e.g., Kaminski, 2016). Teachers commonly express strong intuitions that songs simply 'work' for diverse educational purposes, including memorisation of concepts or vocabulary, improving pronunciation, laying the foundations of grammatical knowledge, supporting classroom routines and behaviour management, 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 outcomes, not just for their motivational power or for teaching music itself. This doctoral research investigates these pedagogical assumptions.

Defining what constitutes 'using songs' to teach languages in the YLL literature is quite slippery. Songs are distinguishable from chants or rhymes, which have salient rhythm but no recognisable melody (Davis & Fan, 2016; Forster, 2006), but sometimes conflated into 'musical activities' including singing songs, reading song books, creating musical instruments, and listening and/or dancing to instrumental music or songs (Paquette & Rieg, 2008). Teachers present songs as individual, small-group or whole-class singing activities, and as listening activities via screen, audio recording, or live performance (Hamilton & Murphy, 2023). Davanellos (1990) proposes 40 activities in which songs are employed, only seven of which involve singing or listening to songs. The rest involve lyrics presented as

gap-filling, sequencing, grammar or vocabulary exercises, or stimuli for creative output (e.g., songs, art, stories). Songs may be presented aurally or in writing, or both (Walker, 2006). The way teachers use the word 'songs' does not always imply that learners are singing in class. In this dissertation, the word 'songs' encapsulates all pedagogical uses that teachers make of songs containing lyrics (i.e., not purely instrumental music). If this does not include singing the song, I will make that clear, otherwise the assumption is that 'using songs' means the learners are *singing songs*. 'Music' will be a broader label that includes songs and instrumental musical activities.

1.1.1 Introducing the study context: songs in England's primary school MFL lessons

Understanding whether songs are an effective FL teaching tool stands to support primary teachers who frequently choose songs for teaching FL because they are perceived as enjoyable and highly effective for achieving linguistic outcomes (Hamilton & Murphy, 2023). If children enjoy learning languages in primary school, *and* make good progress as a result, this might slow the decline in take-up of languages at GCSE, A-level and Higher Education. However, as Macaro (2008) pointed out, introducing FL into the primary curriculum for one hour per week would do little to reverse the decline. According to Macaro (2008), the UK government had not invested enough in primary language teacher education, and providing only one teaching hour per week risked pupils finding FL boring, repetitive, and difficult. He noted a lack of continuity from primary to secondary levels, and too much reliance on the flawed 'the younger the better' trope that implies drip-feeding FL in primary schools would encourage more students to study FL to GCSE and beyond.

Nevertheless, in September 2014, education in ancient or modern foreign languages became a statutory entitlement for key stage 2 (KS2) pupils (aged 7–11 years) in England's primary schools (DfE, 2013a). Due to the lack of primary teacher education and language specialisation, KS2 teachers find themselves responsible for teaching languages they do not necessarily know beyond the basics of what they learned in school (Graham, Courtney,

Marinis, & Tonkyn, 2017) for 30–60 minutes per week on average (Holmes & Myles, 2019). There is also evidence from national surveys (Collen & Duff, 2024; Tinsley & Doležal, 2018), the Research in Primary Languages Summit (RiPL; Holmes & Myles, 2019), and smaller qualitative studies of teachers' experiences (e.g., Finch, Theakston, & Serratrice, 2020) that FL is low in the primary school subjects' hierarchy, side-lined by core subjects, and first to be cancelled when another priority arises. Music faces a similarly precarious position in England's primary curriculum (Fautley, Kinsella, & Whittaker, 2018), which is relevant to how songs are used in primary FL, as explored below.

A growing number of countries are introducing languages early (Cameron, 2001), assuming that learners will pick up languages apparently effortlessly when they are younger (Murphy, 2014). Murphy (2014) expresses concern that YLLs in input-limited classroom contexts are unlikely to achieve the same linguistic outcomes and milestones as YLLs in naturalistic, input-rich contexts. Folk wisdom about an 'age advantage' in starting instructed FL learning earlier has been contradicted in studies of young learners of French in England (Myles, 2017), and of English in Spain (Muñoz, 2006) and Germany (Jaekel, Schurig, Florian, & Ritter, 2017). In each of these longitudinal studies, starting FL instruction later was found to lead more rapidly towards language proficiency than starting before age 11. Given the lack of time spent in KS2 FL lessons and consequent impoverished input experienced by YLLs in England (Holmes & Myles, 2019), songs could perhaps provide a rich source of language input. Indeed, songs are mentioned in the FL primary curriculum as part of exploring the "patterns and sounds of language" to link with the "spelling, sound and meaning of words" as well as part of appreciating the "stories, songs, poems and rhymes in the language" (DfE, 2013a). No guidance is given, however, on how to explore the "patterns and sounds of language" through songs, nor any evidence supplied in support of the statement. These curricular issues are explored in more detail in section 2.2.1.

Many teachers lack specialist FL training and dedicated curriculum time and may believe that younger is better when it comes to FL learning, despite evidence to the contrary. This combination of circumstances leaves primary teachers seeking resources that facilitate what they perceive as a communicative approach (commonly misinterpreted by policy makers as 'target-language-only' instruction (Macaro, 2008)) in the languages they are expected to teach (Garton et al., 2011), and that make lessons fun and enjoyable to avoid pupils perceiving languages as boring, repetitive, or too difficult. It is unsurprising that teachers, tasked with helping students make "substantial progress in one language" (DfE, 2013a) with little training or consistency in provision from one school to another (Finch et al., 2020), draw together resources and tools that support folk beliefs about good language teaching despite little evidence to support their choice (Hamilton & Murphy, 2023).

Primary pupils' motivation for FL learning is generally high (Lanvers, 2017), perhaps because they perceive FL lessons to be 'idiosyncratic' (Finch et al., 2020) in the way they are taught, with more fun and games compared to core subjects. FL is one of the few lessons where pupils regularly sing songs in KS2 (Hamilton & Murphy, 2023). Songs are a freely available 'folk pedagogy' that appears to fulfil multiple educational requirements across affective, cognitive, and academic domains (Forster, 2006; Hamilton & Murphy, 2023; Paquette & Rieg, 2008). Songs are therefore unsurprisingly popular with KS2 languages teachers (Hamilton & Murphy, 2023) as well as international teachers of English YLLs (Garton et al., 2011). Furthermore, whereas music is waning as a well-supported curricular subject in England's primary schools, singing appears to be chosen frequently by teachers seeking to use music in a cross-curricular manner to support children's developing literacy, maths or language skills (Fautley et al., 2018).

The conditions align for songs to appear to simultaneously fill gaps in primary music provision (Fautley et al., 2018), FL provision and training (Finch et al., 2020; Holmes & Myles, 2019), and to help meet primary literacy and maths targets (Hamilton & Murphy,

2023; Lonie, 2010). The question remains, however, as to the extent to which evidence reliably informs teachers about the effects that singing songs has on the linguistic outcomes that matter to them.

1.2 Aims

Given the general scarcity of research carried out with YLLs (Macaro, 2008), and more specifically the lack of reliable evidence to assess whether persistent beliefs that songs confer language-learning benefits for YLLs are true, this study aims to provide a robust and dispassionate assessment of existing evidence and then design a rigorous intervention study to address any gaps exposed by that assessment. Phase 1 of this research, therefore, systematically gathers, appraises, and synthesises the existing evidence about songs' effectiveness as language-learning tools in second or foreign language classroom contexts with YLLs. Building on the systematic review, Phase 2 consists of a randomised controlled trial to empirically assess the extent to which songs contribute to linguistic outcomes of YLLs in taught FL contexts. The overarching aim for Phase 2's intervention study is to provide trustworthy evidence on the effects of songs on the linguistic outcome measured, namely participants' ability to complete the elicited imitation task where they perceived and repeated French sentences presented orally.

1.2.1 Epistemological orientation

There are numerous avenues for researchers to explore regarding the educational and, specifically, the linguistic outcomes of YLLs who receive input in their target FL through the use of songs as classroom activities. A variety of designs, following a variety of epistemological traditions, have been employed by researchers in this field. 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 causal relationships between singing songs and linguistic outcomes, as the following section explains.

This research sought 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. Since there is a clear belief among teachers that using songs is beneficial for improving FL learning outcomes, and specifically linguistic outcomes, the question of causality needs to be addressed. This investigation situates causality within the epistemological tradition of experimentation in social sciences research laid out by Campbell (1957), who stated that "the very minimum of useful scientific information involves at least one formal comparison and therefore at least two careful observations" (Campbell, 1957: 298). Taking this philosophical standpoint aligns this research primarily 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, Cook & Campbell, 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, Heneghan, & 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. All of these designs have been used at one time or another in investigating causal relationships between using songs and FL learning outcomes.

It was the aim of this doctoral research to assess the state of the knowledge through Phase 1's systematic review and then design and conduct as robust an experiment as possible within an educational context (as opposed to a lab-based study) in Phase 2. With the epistemological stance outlined above in mind, the rest of this dissertation will refer to 'reliable' or 'trustworthy' evidence of whether songs 'work' for achieving linguistic outcomes with YLLs as having been established through a fair comparison of songs as a teaching tool

with an alternative or control condition, and with a valid and reliable outcome measure, through a research design that permits attribution of potential effects to the intervention itself and minimises confounding factors. That is not to say that other types of research are not valid or interesting, nor that teachers' personal experience of teaching and reflection upon what works in their classrooms (and years of trial and error with their teaching materials) is not a good indication of where potential effects might lie, but that attributing effects to an intervention requires a particular type of comparative research design to minimise possible biases in the research process.

Given the strong claims I encountered as a practising teacher that songs constitute effective (and even superlatively effective) tools for language teaching and learning, which were echoed in papers written by, with and/or for teachers (Davanellos, 1999; Forster, 2006; Hamilton & Murphy, 2023; Paquette & Rieg, 2008; Saricoban & Metin, 2000; Schoepp, 2001; Walker, 2006), I adopted this rigorous and dispassionate stance to try and establish what we can reliably conclude about using songs with YLLs to achieve linguistic (not affective) outcomes, and to contribute reliably generated evidence to the ongoing discussion.

1.3 Thesis chapter outlines

Chapter 1 of this thesis introduces the study, describing the background to the substantive problem, the outline of the study design, the motivation and rationale, and the educational context in which it takes place. It introduces the aims of the study and the contribution it makes to our knowledge of the topic.

Chapter 2 provides a review of the relevant literature that motivates Phase 1 of the study: the systematic review of intervention studies investigating the substantive linguistic effects of using songs to teach young learners second or foreign languages. Chapter 2 begins by reviewing why teachers might use songs for teaching languages, looking at some of the theoretical motivations for doing so that teachers are likely to have read about in blogs or textbooks, or teacher-facing journals, or that they just feel intuitively. It then reviews the

prosodic bootstrapping hypothesis, a theory of L1 acquisition that is more empirically robust but less talked about in SLA research than the theories introduced in the first section of the chapter. The chapter then narratively reviews existing pedagogical evidence that using songs helps to achieve YLL linguistic outcomes. It finishes by examining evidence from transdisciplinary sources, namely studies about infant pre-speech productions and infant-directed speech and singing that also partially motivate this thesis.

Chapter 3 reports on Phase 1, a systematic review of the intervention literature investigating the substantive linguistic effects of using songs to teach second or foreign languages to young learners in formal educational contexts (i.e., in schools, rather than playgroups or homeschool contexts). The chapter outlines the methods used to gather the included studies and analyse them, and presents the findings as a narrative synthesis, with a quality appraisal of the body of evidence. It finishes by outlining how the systematic review motivates Phase 2's intervention study and the warrant for a randomised controlled trial design.

Chapter 4 outlines the research design and methodology of Phase 2's intervention study, including the data collection and analysis processes. It introduces the context and participants, the materials used to teach the intervention, the screening variables and outcome measure, the data analysis plan (which includes a cumulative link mixed model and Bayes factor calculations) and ethics considerations.

Chapter 5 reports the findings of Phase 2's intervention study. Each of the four research questions from Phase 2 are addressed in turn, with the descriptive findings, the cumulative link mixed model results, and the Bayes factor calculations (for RQs 2a, 2b and 2c). The chapter concludes by summarising the overall findings of the intervention study.

Chapter 6 recaps the research design and methods used in this dissertation and summarises the findings. It then discusses the findings in the light of existing literature and presents implications for teaching practice. It states the limitations of the study and suggests

directions for future research on the topic. Chapter 7 provides concluding remarks on this doctoral project.

1.4 Contributions to knowledge

Phase 1 moves the field forward by addressing longstanding positive bias (in the absence of reliable supporting evidence) towards believing songs constitute effective second language acquisition (SLA) pedagogy. Rather than circulating 'folk theories' about why songs are effective, this study systematically gathers available empirical evidence and presents it with a thorough quality appraisal of the assembled literature, leading to a summary of the overall weight of evidence. To the best of my knowledge, this is the first systematic review and quality appraisal of this area of enquiry. It is also a multilingual systematic review, gathering evidence through searching French, German and Spanish databases, as well as searching in English.

Phase 2 makes a novel contribution to the body of empirical evidence by investigating the effects of using songs on YLL linguistic outcomes, using robust methods and clear substantive outcomes. This is an under-researched area. This study will contribute to our understanding of using songs with YLLs in taught contexts in England's primary schools, a niche that has received little research attention (Macaro, 2008).

Chapter 2

Literature Review

2.1 Introduction

In this chapter I review relevant literature that motivates the first phase of the study: the systematic review of intervention literature investigating the substantive linguistic effects of using songs to teach young learners second or foreign languages. This chapter provides historical, theoretical, empirical, pedagogical and transdisciplinary perspectives on the issue of using songs for teaching YLLs.

Section 2.1 provides a brief history of using songs to teach languages in England to situate the study, and then 2.2 looks at the context of this study in more detail: namely England's primary school FL lessons. I explain the national curriculum aims for teaching FL in key stage 2 (Year 3 to Year 6 (7–11-year-olds)), and songs' place in the 2014 curriculum guidelines (DfE, 2013a). I explore the use of songs as a potential pedagogical tool in primary FL contexts in England, including ways in which teachers use them and share 'good practice' with regard to using songs for FL lessons. Then I examine how current approaches to teaching FL in primary school (e.g., communicative or content-based approaches) may or may not include songs, and highlight any gaps in research regarding best practices or effectiveness.

Following this contextualisation, section 2.3 provides the theoretical heart of the chapter, laying out interdisciplinary evidence to justify my research focus. I begin with theoretical and observational evidence, including the much-cited theories of involuntary mental rehearsal, musical intelligence, and musical learning styles. I then present the prosodic bootstrapping hypothesis, a well-evidenced theory of L1 acquisition, and examine how prosodic patterns (e.g., intonation, stress, rhythm) may help YLLs process and

reproduce new language. I then narratively examine evidence from classroom contexts for L1 and L2 learning.

I finish by examining potential links between singing and language acquisition from transdisciplinary sources, namely studies of infant pre-speech and their cry melodies, and infant-directed speech and singing. The chapter concludes by justifying the need for a systematic review of the evidence from intervention studies as the next step in the research process.

2.1.1 A brief history of the use of songs in language education in England

Pre-20th century

Songs have been widely used in language education in England and Europe since the Middle Ages, when *scholae cantorum* (song schools) attached to monasteries or cathedrals formed a predominantly oral/aural introduction for young oblates[3] to the study of Latin through chanting liturgical music (Murphy, 1968). Boys' vocational preparation for becoming monks began with learning the sung responses of the liturgy, before their more advanced study of Latin grammar and literacy began (Kelly, 1976; Murphy, 1968; Orme, 2006). Such study included (from the 12th century) rhyme, meter and prose as well as syntax (Luscombe, 2004). According to Kelly (1976), sacred music was considered important for teaching Latin, and medieval essays about music focused on pronunciation, particularly the vowel quality, length of syllables and patterns of intonation. When dealing with silent letters, liturgical music tried to guide pronunciation by adding a 'liquescent' note. Thus, a consonant or diphthong that might have been silent in contemporary pronunciation habits was pressed back into existence by being given its own musical note[4]. This focus on pronunciation meant that oblates knew how to pronounce the Latin songs, but not necessarily what the form-meaning links were: in other words, they only had a general idea of what they were singing

[3] Oblates: boys offered by their parents to the monastery for training (Shippey, 2007).
[4] Kelly (1976) remarks that such pedantry restored many letters that had been silent even in Cicero's time.

about until they reached the stage of performing the rituals at Mass, then later added Latin literacy to their phonological knowledge at grammar school.

Shippey (2007) remarks that this phenomenon of knowing the phonological word forms but not the grammar was a typical outcome of such a song-based approach, as satirically captured by Chaucer in *The Prioress's Tale* when a young boy asks his older schoolmate what the *Alma Redemptoris Mater* is about. The older boy replies that it is sung in praise of the Virgin Mary, but "I kan namoore expounde in this mateere, / I lerne song, I kan but smal grammere"[5]. The fact that the phenomenon is immortalised in satirical poetry suggests it was a commonplace experience for boys to recite their Latin songs without knowing what the words meant. Murphy (1968: 6) suggests this lack of deeper understanding is because the needs of the Church fluctuated: sometimes there was "provision for the mere reading of Latin (not necessarily with understanding) and for singing, as well as for more advanced instruction in grammar schools."

Similarly in France, singing prayers (the *Ave Maria*, *Pater noster*, *Salve Regina*) formed the basis of literacy education, with children able to sing in Latin and apparently follow along a text in the primers, despite not having access to the phoneme-grapheme links and even, in the case of Jesuit schools, being denied access to written religious texts in preference of learning the Word by ear to promote holiness and good discipline (van Orden, 2006). Schools existed to furnish the Church with future priests and disciplined congregations (Murphy, 1968), and language learning was primarily a means to advance piety rather than literacy or communicative competence. From such accounts of learning Latin, songs appear to be used pedagogically to promote memorisation of phonological forms, accurate intonation, and a foundation in the sounds of the language, upon which

[5] Translation: I cannot expound on the matter, I learn song but know little grammar.

grammatical understanding might be built, especially if study continued into the grammar school and beyond that to university.

Communicative competence seems to have become more important as French and Italian education expanded in England during the 14th century (Gallagher, 2019). Secular conversational primers such as the *Manières de langage* ('ways of communicating') contained songs intended as pedagogical tools to aid development of oral competence (Leach, 2005). Oral recitation of both poetry and songs in language education continued into the late 18th century, when the grammar-translation method became more popular than reciting verses from memory and composing new ones (Kelly, 1976). The practice of reciting memorised text aloud resurged with the 'direct method' of communicative language teaching that arose in the early 20th century, pioneered by Berlitz. However, this renewed focus on oral practice had the instrumental aim of learning intonation and mastering formulaic expressions for conversation, rather than the more creative and aesthetic aims of the Renaissance and Middle Ages (Kelly, 1976).

20th century

In the 20th century, interest in cultural content reintroduced music into FL classrooms through folk songs and traditional songs, some arranged especially for school use. Where folk songs were seen as archaic, popular songs also found their way into classrooms. Kelly (1976: 100) states that "music and rhythm are extremely important as learning activities" for young children because they are "natural ways in which children learn" and hence became popular in the mid-20th century (Léopold, Jones, Ervin-Tripp, Rivers & Malherbe, 1969). With such a long history of using songs in the education of YLLs, it is unsurprising that the presence of songs is assumed to be of benefit today because it has centuries of pedagogical history in language education.

Whilst songs are certainly part of the rich tapestry of linguistic input that is available for teachers to introduce YLLs to the target FL, and traditional songs form part of the cultures under study, little is known about how songs compare to other teaching resources when it comes to actually learning and making progress in the target language. Anecdotally, many adults have told me that they know all the words to *Frère Jacques*, which they learned decades ago in school, but they do not know what the words mean. Like in the Latin song schools, then, perhaps singing cannot be assumed to teach language to the point of bringing a learner into competent communicative use of the FL, or grammatical understanding of it.

It is also important to note that the purpose of education has drifted far from its ecclesiastical origins (as traced by Murphy, 1968). The monastic and cathedral schools of the Middle Ages had a common purpose of introducing oblates to liturgy – teaching them to read and sing for their vocational purpose (Luscombe, 2004; Orme, 2006) with a "restricted and predictable" curriculum taught by masters appointed by the Church (Murphy, 1968: 5) – whereas today's schools in England follow a curriculum that "prepares pupils…for the opportunities, responsibilities and experiences of later life" (Education Act, 2002; DfE, 2013a), which are diverse and not solely focussed on religion. Historical practices of using songs for learning languages cannot therefore be unquestioningly adopted today with the same benefits assumed as in the Middle Ages. We need to understand the strengths and limitations of using songs compared to other available options in today's FL classrooms in order to use them judiciously and purposefully.

Since the 1960s, there has been a gradually more regulated and centralised curriculum for teaching FL in England, with a prevailing aim of trying to raise standards and students' motivation for learning FL at secondary level (see timelines for summary of key policy developments, Smith et al., 2021; Seccombe, 2021). The progress made in the 1960s towards sustainable FL teaching at primary school level was effectively stalled by Burstall, Jamieson, Cohen and Hargreaves (1974), a "critical and somewhat pessimistic" (Hunt,

Barnes, Powell, Lindsay & Muijs, 2005: 10) or indeed a "kiss of death" (Sharpe, 1989; 1991) report that claimed to find little advantage on long-term achievement of starting French lessons at age 8 compared to pupils beginning at age 11, in spite of considerable methodological flaws that call this conclusion into question (Gamble & Smalley, 1975; Sharpe, 1989; Tellier, 2019). Whilst their pessimistic conclusions were arguably ill-founded, Burstall et al. (1974) noted key factors that impeded the success of introducing FL in primary schools, including inadequate teacher training, unequal provision across primary schools, and poor liaison and continuity between primary and secondary schools: a refrain that echoes through the following decades up to the present day.

In most EU countries, it has been compulsory for pupils to learn at least one foreign language from age 8 to 18 since 2003 (European Commission, 2017). Since 2000, successive UK governments have introduced initiatives to incorporate foreign languages into primary school curricula in England. The National Languages Strategy (*Languages for All: Languages for Life*, DfES, 2002) pledged to put measures in place to give all children in KS2 in England an 'entitlement' to learn FL by 2010. Eight years was a tight timeframe. The strategy recognised that it would be a challenge to implement, both culturally and in terms of planning the curriculum and "mobilising a workforce that has the skills, expertise and confidence to deliver language learning in our primary schools" (DFeS, 2002: 15). Foundations were to be built from existing 'good practice' that was going to be collected from the Early Language Learning (ELL) pilots, funded by DfES from 1999–2001 and managed by CILT – the National Centre for Languages that was formed from the merger of the Centre for Information on Language Teaching and Research (founded in 1966) and the Languages National Training Organisation (founded in 1998).

The National Languages Strategy was followed by the *KS2 Framework for Languages* (DfES, 2005a). The Framework consists of over 100 pages of guidance on developing approaches to teaching FL, with definitions of the expectations, outcomes and

learning objectives for each year group (Y3–6), and suggestions for teaching activities. The five learning objectives centre on: (1) oracy – defined as "listening, speaking and spoken interaction" (DfES, 2005a: 7) – which underpins and is reinforced by (2) literacy; (3) intercultural understanding; (4) knowledge about language (metalinguistic awareness); and (5) language learning strategies (metacognitive awareness transferable to any language). The strands all interrelate, can be developed by schools for their specific context, and elements from across the five objectives are integrated in lessons rather than being taught in isolation.

Songs are mentioned 67 times in the KS2 Framework, primarily as a source of the sound patterns in the FL, knowledge of which will support oracy through helping learners recognise and produce new sounds. The KS2 Framework suggests that the following oracy objectives can all be met through listening to and singing along with songs in the target language: identifying rhyming words, performing finger rhymes and songs, identifying phonemes which are the same as/different from English and other known languages, speaking clearly and confidently, repeating words and phrases modelled by the teacher, remembering a sequence of spoken words, and using physical response, mime and gesture. Since oracy and literacy are mutually supportive objectives, the literacy objectives that make links between phonemes, rhymes and spellings and reading familiar words, poems and rhymes aloud may also be facilitated through songs. The 'knowledge about language' objective includes identifying sounds, phonemes and words, recognising rhyming sounds, imitating pronunciation of words and recognising question forms or negatives, which could again be explored through songs. The 'language learning strategies' objective includes using actions and rhymes to aid memorisation, and remembering rhyming words, both of which link to songs. Indeed, many of the teaching activities for oracy and literacy suggest using songs or (if songs are not specifically mentioned) closely linked activities such as clapping rhythms or focusing on the rhythm of the words in sentences, reading familiar rhymes or poems aloud in chorus, and holding up vocabulary cards when a familiar word is mentioned

in a song or poem. Songs are a thread through the KS2 Framework document, including in intercultural understanding and learning strategies (e.g., 'How do you recall vocabulary in order to sing a song?' for Y6 pupils (DfES, 2005a: 85)). From this framework, the place of songs as an effective and essential part of FL teaching and learning might be assumed. No research is cited to support the framework's suggestions.

Two studies conducted in the 2000s found that the KS2 Framework supported schools in planning their FL teaching. According to a longitudinal survey of primary FL implementation from 2006–2008 (Wade, Marshall & O'Donnell, 2009), 92% of the 500 schools in the representative nationwide sample were teaching FL during class time in KS2, and the KS2 Framework formed the basis of the majority of participating schools' FL programmes. Cable, Driscoll, Mitchell, Sing, Cremin, Earl, Eyres, Holmes, Martin and Heins (2012) conducted a three-year longitudinal (mostly qualitative) study of 40 English primary schools who were 'early adopters' of teaching FL. They reported that participating schools increasingly used the KS2 Framework for planning schemes of work and lessons, focusing primarily on the oracy objective and then literacy. Cable et al. (2012) noted that over the three years, schools tended to offer 30–40 minutes of discrete FL teaching, with a largely oracy-based approach of songs, rhymes, chants, games, and role-play activities aiming to enthuse and engage children.

These two longitudinal studies suggest there was success in increasing primary FL teaching in the wake of the Languages Strategy and investment from the government, with French taught in 89% of schools, followed by Spanish and then German (Wade et al., 2009). From 2002 (the launch of the Languages Strategy) to 2005 (the launch of the Languages Ladder for recognition of pupil achievements and progress), the government invested £22m in MFL education in England and pledged another £115m of investment for the period 2005–2008 (DFeS, 2005b). £137m in total. Amongst other initiatives that built on the Languages Strategy in secondary education, this budget was spent on 1000 primary schools

in 19 local 'Pathfinder' authorities who trialled approaches to teaching languages, plus teacher training. 1200 new and 700 existing primary teachers received specialist MFL training from 2002–2005, with £60m allotted to further training initiatives for 24,000 more teachers. The Languages Ladder had separate qualifications for listening, speaking, reading and writing that could be used with 23 languages for learners at any stage, from Breakthrough (beginners) to Mastery (postgraduate linguists or native speakers). It was an ambitious plan to reverse the national decline in FL skills, although it was somewhat perversely coupled with the government making FL study voluntary from age 14 onwards (effective from 2004). This, critics pointed out, sent confusing messages about the importance of languages study (see for example Guardian, 2002) and seemingly contradicted the *Languages for All: Languages for Life* slogan of the Languages Strategy. Education Secretary Charles Clarke's response to the criticism of this apparent duality was that "children learn best at foreign languages when they're young[...] We've got to strike when the iron really is hot, which is when young people are young at primary school" (Guardian, 2002). Thus, introducing languages earlier was the government's solution in 2002 for reversing the decline in language take-up at GCSE and A Level.

A significant barrier to seeing whether this plan would work over the long term materialised when, in 2011, the newly-elected coalition government discontinued the National Languages Strategy and stopped funding most languages-specific initiatives in schools. At this time, FL study at GCSE was showing signs of dramatic decline: in 2011, 43% of the GCSE (age 16) cohort sat FL exams, down from 78% in 2001 before languages were made optional after age 14 in 2004 (Tinsley, 2013). Conversely, in primary schools there had been an increase in FL teaching from 25% in the early 2000s to 56% in 2007 and 92% in 2010, due to the expectation that FL would become a statutory subject (Cable et al., 2012; Tinsley, 2013). Once again stalling the progress made at primary level, the coalition government spent two years reviewing the national curriculum. Then, forty years after the

Burstall et al. (1974) report that stalled introducing FL at primary level, and despite not having resolved the perennial issues of continuity, progression, timetabling and teacher training (DfE, 2012), the government introduced the statutory study of FL into primary schools in the new national curriculum (DfE, 2013a).

The 2012 DfE consultation on making primary languages compulsory (a survey with 318 responses, 79 from primary schools) contains repeated concerns from respondents about teacher training and confidence to deliver an FL curriculum and how they would get "the basics right…accent, pronunciation and grammar" (DfE, 2012: 7). 20 respondents (13%) pointed out that the *KS2 Framework for Languages* (DfES, 2005a) "was an excellent starting point for practitioners and that this should be re-instated. They believed that teachers had been using the Framework to good effect; and the language learning strategies and knowledge about languages elements in this must not be forgotten" (DfE, 2012: 8). The coalition government did not heed these respondents, and removed the *KS2 Framework for Languages* (DfES, 2005a) from circulation.

In summary, songs have been used in England's FL curricula for centuries, although the purpose of education has shifted from ecclesiastical training in Latin to preparing pupils for diverse life opportunities with communicative competence in a broader range of mostly European languages, predominantly French. Since the 1960s, FL teaching in England has been increasingly centralised to improve standards and motivation for learning languages. After a long hiatus apparently triggered by Burstall et al. (1974), the National Languages Strategy (DfES, 2002) aimed to give all KS2 children the entitlement to learn a FL by 2010. Then the *KS2 Framework for Languages* (DfES, 2005a) defined clear learning objectives for oracy, literacy, intercultural understanding, metalinguistic awareness, and language learning strategies. Songs were central to the oracy and literacy objectives, with songs and rhythm identified as providing support for learning across all five objectives. By 2008, 92% of primary schools taught MFL (89% teaching French), with many using the KS2 Framework

(Cable et al., 2012; DfE, 2012; Tinsley, 2013). Despite significant investment and time spent on the National Languages Strategy, it was folded in 2011 in favour of a new plan created by the coalition government, in consultation with stakeholders, in 2012.

Through the ages, songs have been considered by researchers, stakeholders and teachers to have importance and relevance in the FL classroom, especially for YLLs' oracy. Nonetheless, their effectiveness in developing learners' communicative competence remains unclear from the sources explored in this section. Their historical use cannot be the sole reason for continued unquestioning acceptance that songs contribute something unique to the language learning process, beyond the mere enjoyment they engender. Indeed, the practice needs careful evaluation to determine songs' effects on different linguistic outcomes and best practice regarding their judicious use with learners in a variety of contexts and age groups. A practice that has been used for centuries must surely enjoy some empirical support, but none is cited in all these curricula documents. The following section explores England's 2014 national curriculum (DfE, 2013a) and the place of songs within it.

2.2 Context: languages in England's 2014 primary school national curriculum

Published in October 2013 and coming into force in September 2014, the current national curriculum for England (DfE, 2013a) introduced the study of ancient or modern foreign languages as a statutory requirement, to be taught in all local-authority-maintained primary schools. Immediately prior to this date, there was no statutory requirement to teach FL in KS1 and KS2. However, there were some non-statutory guidelines for primaries wishing to teach FL, drawn from the programme of study for FL at KS3 and KS4 (DfE, 2011). In these 2011 guidelines, songs appear under the "opportunities for the reinforcement of knowledge, skills and understanding developed in other curriculum areas," which could be "exploited through…songs, alphabet, poems, rhymes and stories in other languages" (DfE, 2011). No guidance is offered about how to exploit these opportunities, leaving it up to primary schools to develop their own programmes of study and resources. There is also no mention of the

KS2 Framework for Languages (DfES, 2005a), despite this having apparently been much used by primary schools for FL planning in the preceding decade (Cable et al., 2012; Tinsley, 2013).

"Considerations" in the 2011 guidelines for schools planning to introduce FL include "the availability of suitably trained teachers", "the amount and frequency of teaching time", "the age at which the language is to be introduced" and "continuity and progression from class to class and from primary to secondary school" (DfE, 2011). These considerations differ very little from concerns highlighted in multiple commentaries about introducing FL teaching in primary schools (e.g., Burstall et al., 1974; Cable et al., 2012; Hunt, 2009; Macaro, 2008; Wade et al., 2009). Hunt (2009: 215) considers the lack of "continuity in the curriculum and progression in learning" in core curriculum subjects as a signal that FL will fare even more poorly:

> If difficulties arise in NC statutory core subjects (English, mathematics and science), it is easy to imagine the challenges facing MFL when pupils transfer to a secondary school from a range of feeder primary schools where 'entitlement' potentially means great diversity in language provision in terms of time allocation, teaching quality (both subject knowledge and pedagogic expertise) and even the language studied.

Here, Hunt specifically mentions 'subject knowledge and pedagogic expertise' as contributing factors to the success of bringing FL into primary schools, an endeavour which successive governments hoped would provide better foundations for secondary school language learning (DfES, 2002, 2005b; DfE 2011). Hunt's factors chime with Cable et al.'s (2012) longitudinal observations of the wide variation in teachers' subject knowledge and confidence teaching FL, and also with Graham et al.'s (2017) investigation of the impact of teaching time and teacher factors such as teacher levels of FL training and proficiency on early language learning outcomes in KS2. The amount of exposure time to the FL and also the quality of that exposure (as determined by teachers' FL proficiency, pedagogical skills and teaching methods) are key factors in YLLs' grammatical and lexical development (Graham et al., 2017). Thus, the pedagogy and teaching methods and FL proficiency of

primary teachers are important factors for achieving the aims of the 2014 national curriculum for primary languages, which are outlined next.

The 2014 national curriculum introduces specific aims for the teaching of ancient or modern languages in key stage 2, from Year 3, when children are age 7–8. The purpose, aims, expectations and content for the new FL statutory requirements for KS2 are laid out in a brief 600-word document (DfE, 2013a), which is a contrast to the extensive *KS2 Framework for Languages* (DfES, 2005a). In the 2014 curriculum, FL's overarching focus is for pupils to make "substantial progress in one language" (DfE, 2013a). This is to be achieved through providing an "appropriate balance of spoken and written language" laying the foundations for "further foreign language teaching at key stage 3" and enabling "pupils to understand and communicate ideas, facts and feelings in speech and writing, focused on familiar and routine matters, using their knowledge of phonology, grammatical structures and vocabulary" (DfE, 2013a). 'Practical communication' is the focus for modern languages, whilst ancient languages focus more on reading comprehension and appreciating classical civilisation.

2.2.1 Songs and the 2014 national curriculum for primary FL

Songs are specifically mentioned twice in the 2014 national curriculum, first as a means to "explore the patterns and sounds of language" and then as teaching the pupils to "appreciate stories, songs, poems and rhymes in the language" (DfE, 2013a). Schools are left to devise their own pedagogy and materials: the curriculum itself provides no guidance on how to use songs to explore the patterns and sounds of language. However, there is a video (DfE, 2013b) of Linda Dupret, headteacher of St Paul's Primary School, Brighton, who encourages teachers to:

> make sure you use a real range of methodology, stories, rhymes, songs – songs are very powerful in learning a new language because the children really enjoyed them and it sticks in their mind.

Dupret singles out songs as 'very powerful', which echoes practitioner beliefs that songs are superlatively effective for teaching languages (Hamilton & Murphy, 2023). Indeed, Dupret's words echo those of a KS2 languages teacher interviewed in Hamilton and Murphy (2023: 1497) that "words just go in your head without thinking about it when you learn through a song." Dupret also explains that her school always has four or five Spanish fluent or native speakers to help immerse the pupils in Spanish, a Spanish exchange school visit where classes become 'bilingual' for the duration of the visit, plus pupils have at least an hour of Spanish lessons every week. These aspirational conditions for learning Spanish are not representative of the average primary school in England, where the lack of teaching time and training in FL are key barriers to creating a rich target language environment for learners (see the *Language Trends* reports, e.g., Collen & Duff, 2024; Tinsley & Doležal, 2018). Her encouragement to teachers to use songs, however, feels more achievable and may therefore be the practical 'takeaway' teachers glean from this video guide.

The lack of guidance in the 2014 curriculum combined with teachers receiving little (if any) specialised training in how to teach FL at primary level leaves something of a vacuum, especially when coupled with the decreasing funding for FL. 80% of schools responding to *Language Trends 2024* had no funding for FL in 2023/24, 10% more than the previous year (Collen & Duff, 2024). The trend is for schools to fill this gap in training, time for planning, and subject expertise by increasingly relying on commercial materials for teaching FL rather than creating them in-house. 79%[6] of primaries who responded to the *Language Trends 2024* survey said that they bought in commercially produced materials (an increase from 75% in 2023) compared to 40% preparing resources in-house (down from 49% in 2023; Collen & Duff, 2024). The list of commercial or external materials given by respondents includes: "Hackney education resources, The Language Gym, Primary

[6] That's 476 of the 603 total responses received. n = 603 represents 10% of primary schools in England.

Languages Network, KAPOW, Light Bulb Languages, local Language Hub, Rachel Hawkes' Scheme of Work, Language Angels, and Twinkl" then "specialist resources that teachers themselves have created" (Collen & Duff, 2024: 9).

Songs are included in several of the commercial and freely available ready-made FL materials providers. KAPOW has a French weather rap, finger rhymes, and lessons on writing a song in French. Primary Languages Network materials include songs that are specially composed to focus on a language feature, with accompanying PowerPoint presentations to pre-teach the vocabulary. For example, the song *J'aime les animaux*, sung to the tune of the Hokey Cokey, focuses on animal nouns and the question *qu'est-ce que tu aimes?* There are songs for teaching classroom commands (e.g., *écoutez*, *regardez*) and manners (*merci*, *s'il vous plaît*), as well as greetings (e.g., *Comment t'appelles-tu?*) and songs linked to festivals (e.g., Easter and Christmas), or topics (e.g., colours). Freely available, the Rachel Hawkes (2024) KS2 schemes of work include songs in the Phonics strand (listening and production) with statements about songs in each year group's learning outcomes from Y3–6, e.g., *I enjoy listening to and joining in with simple songs and rhymes* for Y3. Also freely available, Light Bulb Languages (2025) provides a selection of traditional French songs (e.g., *Frère Jacques*) and a rationale for including rhymes and songs in KS3, stating they would "typify current practice" in KS2 language teaching:

> Giving language a rhythm and a tune often makes it "stick" better, and enables you to tackle much longer passages of the language than you might attempt otherwise. The language used is necessarily repetitive but not boring, and therefore more motivating. Learning through song accesses different learning styles and enables students to interact with the language in a different way, within a dynamic and social environment. It also provides a safer environment in which to explore and experiment with new sounds, with students' participation increasing as their confidence grows. Children are naturally interested in music and songs anyway, so it seems a sensible thing to include in MFL lessons in all key stages. If you can find songs from other countries where your language is spoken, it's an easy way in to Intercultural Understanding, something that Ofsted says no key stage does enough of.
>
> (Light Bulb Languages, 2025)

This rationale claims several unsubstantiated benefits of using songs for teaching FL in KS2/3, and provides two (now defunct) links to background reading to support them. The idea that songs 'stick' is an echo of Dupret (DfE, 2013b) and Hamilton and Murphy (2023), discussed above. Overall, however, there is little guidance on what to do with songs or how they are useful for learning languages in any of the listed resources, or in the national curriculum itself. The following section explores how teachers use songs in FL lessons.

2.2.2 How teachers use songs in primary FL in England

The previous sections explored how songs are included in the national curriculum (DfE, 2013a) and in commercially or freely available schemes of work for KS2. This section looks at how teachers use songs, how they feel about the available materials that guide them, and what is considered to be 'best practice'.

An area where teachers find songs useful relates to the scant available curriculum time for FL in primaries, noted in successive *Language Trends* reports (summarised in Collen & Duff, 2024), longitudinal studies (Cable et al., 2012), and the RiPL white paper (Holmes & Myles, 2019). Given that the "jam packed National Curriculum" (Collen & Duff, 2024) does not give many primary schools time to dedicate more than 30–40 minutes to languages each week (Cable et al., 2012; Collen & Duff, 2024; Holmes & Myles, 2019), singing is one way (as well as themed assemblies, performances and projects) that teachers provide extra time for language learning, without taking time away from other subjects (Tinsley & Doležal, 2018). Singing in the FL is thus a cross-curricular activity that meets several objectives simultaneously (such as doing a drama performance with singing in the FL) and helps with the challenge of finding sufficient curriculum time for languages.

Another challenge for successful language teaching in primaries is "staff confidence and expertise" in languages (Collen & Duff, 2024: 12). Reliance on commercial materials for planning and teaching FL is increasing (Collen & Duff, 2024). There is evidence that teachers (especially if they are not confident at speaking the language they are teaching) rely

on commercial materials to present the FL orally for them. On a Reddit social media thread discussing the merits of commercial FL schemes (Axehandle1234, 2022), one anonymous primary teacher explains that their "current school buys in a scheme and teachers just muddle through it" (LostTheGameOfThrones, 2022). Another writes:

> I teach French to my key stage 2 class and I am terrible at it, to be honest. We use the scheme *Language Angels* and it feels me with dread each week! The scheme is quite good though – this is my first time teaching it as I've given French to my PPA cover the last few years (as most people have, to be honest...) and it literally says everything for me, has songs, games etc. So it's better than before but I know it's not good, and would only really be better with a proper specialist teacher.
>
> (Unopeia, 2022).

These are anecdotal examples, yet it is illuminating to think about how much skill is required to even 'muddle through' a bought-in scheme of work if teachers have little spoken FL proficiency or confidence speaking the target language themselves (Collen & Duff, 2024; Graham et al., 2017). It is understandable, therefore, that Unopeia (2022) seems relieved that their scheme "literally says everything for [them], has songs, games etc" as they dread teaching French and feel terrible at it, but still recognise that "it's not good" and requires a "proper specialist" to make the FL provision better. One use, therefore, that teachers are making of songs is to provide authentic FL content, with correct pronunciation, to do the speaking for them, since this is a key component of helping pupils "explore the patterns and sounds of language" (DfE, 2013a). It does not appear that the teacher is always singing the song with the class, but presenting it as a video or sound file embedded within the scheme of work.

Indeed, with music also being increasingly squeezed in curriculum time at primary level (Fautley et al., 2018), FL is one of the few lessons where pupils regularly sing songs in KS2 (Hamilton & Murphy, 2023). Section 1.1 explored how teachers of young learners in the UK and internationally use songs for achieving diverse educational purposes such as

introducing content, promoting positive affect, supporting behaviour management and signalling classroom routines (Hamilton & Murphy, 2023) as well as for achieving linguistic progress in vocabulary, grammar, and pronunciation (Paquette & Rieg, 2008; Forster, 2006; Walker, 2006). In the context of England's primary and secondary school languages teachers, there is further evidence that songs are a valued part of classroom routines and languages lessons. Several websites and blogs from experienced MFL teachers sharing ideas for practice mention the benefits of songs. Smith's (n.d) reasons why teachers might use songs align with Hamilton and Murphy's (2023) findings: music lyrics are a source of comprehensible input and culture from the FL-speaking countries (content); songs calm down and help control students (behaviour); songs are relaxing and help students feel comfortable to use the FL (social/affective); and they are memorable and fun.

One of the difficulties enumerated in Degrave (2019) for using songs for language teaching is finding suitable song materials that are age and stage appropriate and contain useful FL content. The teaching blog *Geraldine Teacher* suggests extracting short clips from songs to create a video montage that contains song clips with pertinent linguistic features and age-appropriate content (Ubeda, 2018). Ubeda (2018) recommends reading Conti (2015), who expounds 'how to exploit the full learning potential of an L2 song in the language classroom' in a blog containing a detailed exploration of using songs for language learning. Conti (2015) begins by explaining why, whilst some incidental learning may occur, he does not believe "lyrics' key vocabulary or structures" will be learned from songs without a "principled approach" to using them. His nine-step framework (containing 14 steps) details the journey language teachers might take during the 'exploitation of a song', from selection, pre-listening activities, listening for pleasure and then listening tasks (step 4), recognising and noticing vocabulary items, focusing on phonemes, segmentation, sounding out, reading comprehension (lexis, grammar and syntax), meaning construction, translation, singing along (step 11), recycling and consolidation, and finally to reflection on how songs might

help students learn the FL. It is a comprehensive framework with all four modalities (listening, reading, writing and speaking) and vocabulary, grammar and phonics all involved in the process of using a song for language learning. It seems like a feasible and useful approach, hence being called a 'must-read' when recommended by Ubeda (2018), but it is aimed at secondary level rather than primary level. Also, there is a lack of accompanying research evidence that might reassure teachers that Conti's approach enjoys empirical support, or help to assess the applicability of these approaches with primary-aged learners.

A standalone multilingual music project in Wales, Cerdd Iaith (Listening to Language), saw teachers at some primary schools teaching songs in Welsh, Spanish and French, in collaboration with linguists and professional musicians. There was a focus on single words first, with each syllable given a musical tone and rhythm to create mini units of music. All three languages were presented together with images and the musical units, which were then combined in different ways (starting from the musical phrase or from one of the languages while children responded with a matching item in one of the other two). More complex units were then introduced, and then songs in all three languages together, culminating in a performance. In a British Council blog about 'why rhyme, repetition and rhythm are so effective in helping us learn a language' (Mordsley, 2017), a teacher from Cerdd Iaith describes the way a pupil retrieved a vocabulary item (yesterday) in Welsh by humming the tunes of the words *yesterday* and then the Spanish *ayer*, before finally arriving at the Welsh *ddoe*. The teacher saw the benefits of "using music to make connections between languages" (Mordsley, 2017). Several more British Council blogs also provide teaching ideas on how to use songs in languages lessons (Míguez, 2017; Simpson, 2015).

Overall, in addition to the ways of using songs mentioned in schemes of work (section 2.2.1), teachers use songs in primary FL as a way of meeting the national curriculum targets despite the key issues of curriculum time, teacher confidence and available FL materials posing a challenge. Songs are used for including languages content

into non-timetabled subjects or routines, to save time by multitasking; as a way of presenting the FL when teachers do not feel confident speaking it themselves; and for exploiting or teaching language content.

2.2.3 Approaches to FL teaching with young children

The above example of Conti (2015) prompts the reflection that materials or schemes aimed at secondary level pupils cannot necessarily be adopted verbatim, nor even perhaps easily adapted, for primary level FL. Approaches to FL teaching with children in primary school quite often stretch out the approach taken in early secondary (Y7), diluting a secondary syllabus by simplifying verb structures to include, for example, only present continuous verbs (Cameron, 2001). However, there is no unified nationwide or DfE-prescribed approach for teaching FL at primary school. The nearest England apparently came to that was the *KS2 Framework for Languages* (DfES, 2005a), which was quite widely adopted by primary settings but abandoned after the government changeover in 2011 (see section 2.1.1). This section examines some of the approaches suggested since the early 2000s in volumes aimed specifically at teaching FL in primary schools, as well as approaches observed through surveys of teaching practice, and describes the place of songs within these approaches.

Firstly, I provide an extremely brief recap of key trends in language teaching approaches to situate the current primary school FL context. Language-learning pedagogy in the last two centuries has cycled through several trends in approaches, from a predominantly Grammar–Translation approach before the 20th century, through Direct Method (spurred by Berlitz), then Audiolingualism (in the 1960s) and humanist approaches such as Total Physical Response (in the 1970s). Since the 1980s a predominantly input-driven communicative approach has dominated (Hymes, 1972; Canale & Swain, 1980) in the EU, with a comparable Natural Approach (Krashen & Terrell, 1983) trending in the US, whereby input provided by teachers leads to production naturally emerging, with little explicit teaching of forms or corrective feedback. The 'strong' form of communicative language

teaching (CLT) involves engaging learners in meaningful FL conversation with no explicit grammar teaching, whereas the 'weak' form of CLT involves structured activities following a typical Presentation–Practice–Production (PPP) format.

From the 1990s, a focus on form (FonF) approach (not focus on formS as in Grammar–Translation and Audiolingualism; Long, 1997) has influenced teaching approaches with a meaning-based approach that is similar to the weaker form of CLT. Starting off with fairly formulaic set phrases, learners progress to more authentic and creative communication in all four skills of speaking, listening, reading and writing. FonF has been found effective for learning (Ellis, 1995; Lightbown & Spada, 2003) but Kirsch (2008) cautions that it is easy to slip into Grammar–Translation and Audiolingual drill-based modes, as will be explored in the following section. Finally, task-based instruction (Ellis, 2003; Long, 2015; Nunan, 2004), which is based on sociocultural theory, involves role-plays and real-life situations with a problem-solving element that involves negotiating meaning through the FL in a culturally appropriate manner in structured activities of pre-task teaching (e.g., of key vocabulary), tasks in pairs or groups, planning of oral and written reports, and then reporting (following the structure proposed by Willis and Willis, 1996). All of these approaches can be seen as partially influencing how FL teaching is approached in England's primary schools, whether that is encouraged through government documents (such as the *KS2 Framework for Languages*, DfES, 2005a) or observed and reported in surveys of teaching practice (Cable et al., 2012; Driscoll et al., 2004; Muijs et al., 2005).

Martin (2000) suggests that there is no single 'right way' to implement primary FL, but there are three broad possible approaches that shape how schools organise FL instruction. The first approach – language competence – focuses on linguistic progression in the FL (probably just one language). This approach requires more dedicated FL time, for teachers to know the target language, and for secondary schools to acknowledge children's prior learning and maintain their progress when they reach Y7. Martin suggests the second

approach – language sensitisation or encounter – is perhaps best suited to England's context where teachers often lack the confidence, training, and time to teach language competence programmes. This 'encounter' approach focuses on developing positive attitudes to language learning, with some basic competence and knowledge of formulaic phrases. The third approach – language awareness (Hawkins, 1984) – contains less FL content and is more a preparation for language learning, with meta-skills such as learning how to learn languages, and awareness of linguistic diversity. This awareness approach requires much less extensive FL knowledge on the part of teachers, and may facilitate transition to a language competence approach at secondary school. Martin (2000) writes that literacy can be promoted as part of the language awareness approach, since learners with a basic knowledge of their L1 literacy can learn to discriminate sounds, link sounds and written forms, and "how to match sound to print by shared reading aloud of familiar texts, using poems, rhymes, songs, stories, and 'big books' in other languages as well as English" (Martin, 2000: 7). Under this approach, therefore, FL teaching and learning would tie in with the core focus on English literacy in primary schools (an aim later evidenced by Murphy, Macaro, Alba and Cipolla, 2015), and songs may have a part to play in that.

Survey and observational data from Powell et al. (2000), Driscoll et al. (2004) and Muijs et al. (2005) identify Martin's (2000) three approaches to teaching FL in practice in primary schools in the early 21st century: the competence, language awareness, and sensitisation approaches. The competence approach was not often observed since it required considerable time and expertise, and the language awareness approach was also not often encountered. The sensitisation approach was the dominant model used in primary schools at this point, perhaps because its broad aims and simpler language content were more suitable for a generalist teacher, and it involved shorter bursts of FL that included songs, games and integrating FL in daily routines with a focus on affective rather than cognitive factors (Kirsch, 2008). Lessons were reportedly often teacher-led with a minimal amount of

vocabulary and structures repeated extensively through intensive oral practice drills of questions and answers, rote learning, and plentiful repetition in songs, games and rhymes (Powell et al., 2000; Muijs et al., 2005). Pupils reported finding FL lessons fun and less hard work than other subjects because they were full of games. This finding is mirrored more recently in Finch et al., (2020) where teachers reported their pupils enjoying FL because the lesson content is idiosyncratic and more fun compared to their core subject lessons.

Oracy has long been a key focus for FL approaches in KS2 (Cable et al., 2012; Muijs et al., 2005), with songs mentioned in tandem with oracy approaches in the KS2 Framework (DfES, 2005a), related schemes of work (QCA, 2007), and teaching guides for primary FL (Cameron, 2001; Forder, Phillips and Watts, 2013; Kirsch, 2008). To avoid lengthening the time primary children are exposed to simpler versions of a secondary syllabus, as this will be repeated in early secondary school and potentially demotivating (see also Macaro, 2008), Cameron (2001: xiii) suggests giving primary children "a broad discourse and lexical syllabus." Cameron (2001) defines discourse as 'language in use' such as a text (like a shopping list) that is written and used by a particular person, conversations, stories or songs, and other larger units of talk that also spring from using language for real purposes. Her "broad discourse" then might include a range of songs that enable learners to "notice the details of how the foreign language works, from the inside of words up to the large units of stories or descriptions" (Cameron, 2001: 242). Cameron goes on to say that learners "need to incorporate this knowledge through use, and to be able to use the knowledge in their own communication" (p.242). Coupled with her exhortation that time for FL is "too short to waste on activities that are fun but do not maximise learning" (p.2), it could be surmised that songs should be chosen for a purpose of helping learners encounter new language and its meanings, and to help bring that language into communicative use, which echoes Conti (2015) in its sentiments that singing a song is not necessarily effective for learning language if only incidental learning is intended.

Following the KS2 Framework (DfES, 2005a), the Qualifications and Curriculum Authority (QCA) proposed one approach for meeting the Framework's objectives in an 'optional' scheme of work (QCA, 2007). It adopts a language competence approach, assuming no less than 60 minutes of dedicated FL time per week. Regarding oracy, the QCA approach acknowledges that the input-limited context of England's primary schools requires that YLLs regularly and frequently hear a good model of FL pronunciation. Songs are included as part of the proposed sequence for teaching that follows a Presentation–Practice–Production format. A variety of presentation approaches are encouraged, including songs. But it is at the practice stage that songs (and other enjoyable activities such as rhymes and games) are particularly mentioned as enabling "children to repeat new language in a motivating way" which will eventually lead to children applying "the new language in a new context by adapting and adding to it" (p.18). Songs to practise greetings and numbers are introduced in the first unit of work, with *Sur le pont d'Avignon* included as a way of YLLs identifying sounds and raising their hand when they hear the '*on*' sound, for example. For production, there are a number of ideas for including songs (or stories, poems and sketches) as end-of-unit production activities in performances in assembly or to the class (QCA, 2007).

In anticipation of FL becoming a statutory requirement in 2014, Forder et al. (2013) propose an 'integrated approach to teaching foreign languages in primary schools' whereby FL learning is structured around the wider curriculum to embed languages into different contexts, such as history, science, PE, art or mathematics. They propose a wide variety of creative activities where songs are included, some organised around festivals. For example, they suggest singing *Mein Hut hat drei Ecken* in German for European Day of Languages, providing instructions for actions and hat-crafting activities to accompany the song. For Mardi Gras, the authors suggest making pancakes and then performing a song in assembly that recycles the recipe instructions (in French) to the tune of *London's Burning*. As well as the festival-linked song activities, there is a whole chapter about using songs for teaching

languages. The rationale given is that "singing is a lively, active way to teach vocabulary in class, and most students love to sing in the safety of a group if it's fun" (Forder et al., 2013). Most of the songs suggested are familiar tunes with lyrics made up of short items of FL vocabulary (e.g., *Un croissant, un croissant | Un chocolat chaud pour moi* sung to the tune of *London's Burning*) or short sentences (e.g., *Je bois du café | Je mange un croissant chaud | Je prends le petit déjeuner | Bon appétit!* to the tune of *Frère Jacques*). This approach does not involve traditional French songs (other than the tune of *Frère Jacques*) for promoting intercultural awareness, but songs that are musical vehicles for recycling and repeating vocabulary and formulaic structures.

Since the melodies in Forder et al. (2013) were not written to accompany French speech patterns, sometimes the number of proposed syllables in a line does not synchronise with the number of notes in the melody. For example, in one proposed song the line *Moi j'ai faim* has three syllables, but the final line of *London's Burning* has four notes. This mismatch raises the question of which word should have two notes accompanying it, and whether this helps or hinders pronunciation or awareness of French speech prosody. In another song, the tune of traditional French song *Sur le pont d'Avignon* is given new lyrics about *Monsieur Jacques'* daily routine and how he eats *frites tous les lundis*. It is also proposed that new vocabulary be introduced using the tune of *London's Burning* and adding more vocabulary with each repetition. However, it would be useful to know whether these musical and integrated approaches help children learn French in a way that promotes communicative competence, or whether they are simply a fun introduction to the sounds of French (or other languages), or whether using traditional French folk songs would be more effective (and for what purposes, specifically). Many questions remain open about the linguistic effects of using songs in the proposed integrated manner, and whether in terms of children's affect, adding actions like thumbs-ups really "make the song more exciting" (Forder et al., 2013: 159) and whether and how this impacts FL development.

Kirsch (2008) has a chapter on how to introduce young children to languages, emphasising fun and enjoyment but stressing that "language learning is more than amusing children with playful activities for five minutes a day. For pupils to acquire a language, lessons need to build upon each other and offer a range of communication situations. This enables learners to do something in the new language and to make progress" (p.81). In the chapter, Kirsch (2008: 85) proposes several benefits of using rhymes, poems and songs. As well as being popular with YLLs, Kirsch writes that they are familiar to teachers, who may "thus find them a good way into the teaching of foreign languages." Kirsch also states that "the rhythmical patterns facilitate and accelerate learning" and "they are a good way of developing listening, pronunciation and speaking skills." This, she says, is because "pupils do not tire of listening to and repeating them over and over again. They join in with the parts they know and acquire more sounds, words and sentences with each successive performance until they gradually master the text." Kirsch also asserts that poems, rhymes and songs "help pupils get into the rhythm of a language and learn to pronounce sounds and words confidently, accurately and with expression." Furthermore, "pupils are more likely to remember the new words and structures because they are repetitive, meaningful and presented in predictable patterns and larger chunks. The internalisation of sounds, words and sentence patterns bring learners a step closer to using these in other contexts," she says.

There are a number of statements made in this list of advantages that would benefit from substantiation by empirical evidence to demonstrate any causal links between rhymes, poems and songs and linguistic outcomes. None is provided. Perhaps Kirsch considers the group experience of "many language teachers" (p.85) that begins the section is sufficient. It would be helpful if the mechanism by which internalising song lyrics is assumed to promote YLLs' use of FL knowledge "in other contexts" were discussed, or even articulated. This transfer from learning a song to productive and communicative use of the FL appears to be an assumed outcome of using songs, rather than an explicit process teachers need to follow,

perhaps influenced by the ideas of the Natural Approach (Krashen & Terrell, 1983) or weak CLT approach.

Regarding the mechanism through which songs (and rhymes and poems) develop YLLs' FL outcomes, the emphasis in Kirsch (2008) is on introducing, recycling and expanding learners' vocabulary through using songs as a way of drilling vocabulary items. Kirsch (2008) notes that "many teachers take advantage of the popularity and repetitive structure of songs to practise key vocabulary in an *enjoyable* way" (p.85; italics in original perhaps indicating a contrast to audiolingual drills) by making up their own lyrics to traditional or well-known melodies.

Kirsch urges teachers to pay attention to "both the number of syllables and the intonation pattern" (p.96) to ensure that pupils learn the correct pronunciation of words. The exemplar songs are all thematic groups of words or short phrases (e.g., greetings and introductions) set to popular tunes such as *She'll Be Coming Round the Mountain* or *Sur le pont d'Avignon*. In two example songs, German words of greeting or introduction are matched with these traditional tunes. The intonation and melody match well with the syllables in the phrases for both songs, which are introduced as "home-made" songs that "teachers can use to teach, practise and revise greetings" or "a nice way to recycle the vocabulary on greetings" (p.88). Using familiar tunes is said to allow pupils to focus on any new language. For example, if they know some German greetings and the tune to *Two Little Dickie Birds*, they can focus on the new phrase *Ich heiße*. In this way, Kirsch (2008) proposes how songs might permit a simple expansion of YLLs' vocabulary, building on familiar foundations.

Songs can also be used to build up knowledge of a semantic area, such as body parts, by exploiting the repetitive and easily-learned verse structure of songs such as *Heads, Shoulders, Knees and Toes* or *Alouette, gentille Alouette* (Kirsch, 2008). The French song *Loup y es-tu?* is popular for learning clothes vocabulary because a new item of clothing can

be added in each verse. Kirsch (2008) also suggests ways of comparing words across languages, such as *Knie* in German and *knee* in English, as the song will provide a useful context for such comparative analysis. Kirsch proposes expanding this song with follow-up work discussing the origins of English and German, the different pronunciations of the silent/pronounced '*k*' in *knee*/*Knie*, and phoneme-grapheme correspondences that could link to the English literacy curriculum. Indeed, from a seemingly simple starting point of a familiar song, there is a substantial amount of age-appropriate language awareness work that could be developed.

Regarding the motivational or affective benefits of using songs for FL teaching, another key aim of beginning FL teaching early (Kirsch, 2008), Driscoll and Frost's (1999) volume on teaching modern languages in the primary school contains a chapter on using games and songs (Rumley, 1999) where it is proposed that learning a song positively improves children's sense of self-efficacy. Rumley (1999) describes the Kent primary MFL project, which assumed teachers were non-specialists and that classroom teachers, being the children's most significant person in the school day, were the best model for having a go at languages. They aimed to exploit a little bit of language, so teachers could recycle a small repertoire but use it in real communicative contexts, building on Sharpe's (1991, 1995) approach of including FL in classroom routines (e.g., greetings, taking the register, taking the lunch orders). Classroom routines present opportunities to use another language for a real purpose but do not take up any curriculum time. Rumley (1999: 125) advocates using songs because "they provide a safe, non-threatening context within which to play with language" and "provide excellent opportunities for repetition and practice which would otherwise be tedious." Rumley (1999) suggests using familiar tunes like *Old MacDonald*, which are readily adaptable, and traditional songs like *Frère Jacques*, which are readily available. The Kent approach apparently builds children's self-efficacy because if children can do the songs and games, they will feel good about learning languages. The approach is less about teaching

expansive FL content, and more about facilitating children's attitudes towards learning languages in readiness for secondary school specialist teaching.

More recently, Jones and Coffey (2016) suggest that core principles of primary FL are cultural learning, differentiation, language awareness and transferable skills, and that age-sensitive approaches to teaching are necessary to transfer effectively to secondary school FL learning. The authors propose a content-based approach to FL instruction with drama, games and creative activities to support whole curriculum learning.

In summary, teaching FL in primary school might entail a language competence, language sensitisation, or language awareness approach (Martin, 2000); a broad discourse and lexical syllabus approach (Cameron, 2001); a whole-curriculum integrated approach (Forder et al., 2013); a loosely-speaking weak CLT or Natural Approach (Kirsch, 2008); an 'activities and routines'-based approach (Rumley, 1999); a content-based and creative activities approach (Jones & Coffey, 2016); or a combination of approaches, influenced by different aspects of FL teaching approaches from repetition through drills (e.g., Audiolingualism) to input-driven CLT and meaning-based FoF. Songs play a central role across these approaches as tools for promoting cultural and phonological awareness; reinforcing vocabulary and rehearsing formulaic phrases; and boosting learner confidence, motivation and self-efficacy, especially in view of fostering positive attitudes towards FL in preparation for secondary level FL instruction. Nonetheless, none of this addresses the important question of whether songs are effective in promoting communicative competence; nor the relative merits of different uses of songs and approaches for achieving linguistic learning outcomes; nor how different types of songs may (or may not) align with FL speech prosody (and whether this is important); nor how songs contribute to helping YLLs notice forms without explicit teaching. The following section looks at selected theories that posit reasons for why songs might constitute effective FL pedagogy for YLLs.

2.3 Why might songs constitute effective FL pedagogy?

The previous section explored how songs are proposed as an effective way of meeting FL teaching objectives, particularly for oracy objectives, through the national curriculum (DfE, 2013a), frameworks and schemes of work (DfES, 2005a; QCA, 2007), observation of teaching practice and self-reported teaching experience (Cable et al., 2012; Collen & Duff, 2024; Hamilton & Murphy, 2023), and teacher-researcher guides to teaching FL (Forder et al., 2013; Kirsch, 2008). This section explores some key theoretical bases for thinking that songs are an effective pedagogy for FL teaching and learning, beginning with theories that are often cited by practitioners (and researchers in classroom studies), and moving on to look at the prosodic bootstrapping hypothesis. Then follows an exploration of evidence derived from L1 and L2 classroom studies. Finally, evidence that songs and singing are influential for language acquisition derived from transdisciplinary studies of infant pre-speech and infant-directed speech and singing are presented.

2.3.1 Theoretical and observational evidence

There appears to be a lack of well-grounded theoretical motivation for research conducted into using songs with YLLs. As explored in section 2.2, policymakers and 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 (i.e., these beliefs become a kind of 'folk theory'). 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 & Gant, 2016). Establishing stronger theoretical motivation for why songs might support FL learning in classrooms would provide more solid foundations for research to build on.

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.

2.3.1.1 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 evidence regarding the din's supposed "real practical value" (Krashen, 1983:44), its proposed psycholinguistic mechanism, or its 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, 1990; 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 that songs that are easier to sing along to (e.g., *Baby Shark*) are 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 their involuntary intrusion into the mind is their key characteristic. 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 exploring the theoretical motivations for their investigations into the effects of songs on linguistic outcomes.

2.3.1.2 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, the paper 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.

2.3.1.3 Prosodic bootstrapping hypothesis: from L1 to L2 theory

Another potential theoretical foundation for empirical research in this field is the 'prosodic bootstrapping hypothesis' (Gleitman & Wanner, 1982; Morgan & Demuth, 1996), whereby infants exploit acoustic cues in the speech signal to assist in acquiring lexical items, grammar and morphosyntax (Nespor & Vogel, 2007). To contextualise the discussion, this section gives a brief overview of three prominent L1 bootstrapping[7] hypotheses (semantic, syntactic and prosodic) and their foundations in nativist or usage-based accounts of L1 acquisition; then outlines in more detail the key principles of the prosodic bootstrapping account of L1 acquisition; and then examines how prosodic bootstrapping may provide a

[7] A bootstrap is the small strap at the heel of a boot that allows the wearer to pull the boot on. The term 'bootstrapping' was first used figuratively in computer science, where early computers were primed ready to load operating systems with an initial smaller 'bootstrap' program. Thus the term bootstrapping refers to how a large and complex system is leveraged by a smaller initial primer or program (Höhle, 2009) and was introduced to applied linguistics by Pinker (1984).

basis for understanding the mechanisms of primary school children's FL development in instructed contexts as well as infants' L1 acquisition in naturalistic contexts. Firstly, I provide a brief introduction to linguistic prosody and how it is related to meter in verse and music (particularly setting text to music), since an understanding of these terms will inform the rest of this thesis.

2.3.1.3.1 What is prosody?

The word prosody comes from the Ancient Greek for 'song with accompaniment' (Matthews, 2014) and traditionally involves the study of formal verse meter, the metrical patterns used in poetry and oration (details in the following section). In the mid-20th century, research began investigating the inherent suprasegmental features of spoken language and communicative functions of prosodic patterns (Hayes, 1989; Lehiste, 1970; Nespor & Vogel, 1986; Pierrehumbert, 1980; Selkirk, 1984). Both meter and spoken prosody involve acoustic properties of language, but prosody is actually multimodal, encompassing visual as well as auditory cues and discourse markers (Hirschberg, Benus, Gravano & Levitan, 2020; Swerts & Krahmer, 2020). Since auditory prosody is the most pertinent to my thesis, I use *prosody* to mean *auditory prosody* unless otherwise indicated.

A standard definition is somewhat hampered by ongoing shifts in conceptions in the field, but broadly speaking, prosody can be defined through the interplay of its function and form, and how its key components of tone, stress, intonation, and prosodic constituents (a 'prosodic hierarchy' from morae and syllables to utterances) contribute to creating communicative effects or meaning separately to, but inextricably woven with, lexical choice (Gussenhoven & Chen, 2020; Liberman & Prince, 1977; Wagner & Watson, 2010). Prosody can be thought of as 'organising' the speech signal. Its function as a meaning-making element of spoken language is unbound from lexical or morphological meaning, instead referring to how items in an utterance are grouped rhythmically, how they relate to each other semantically and syntactically, the type of speech act being uttered, the speaker's emotions

or attitude, and where they place emphasis (Wagner & Watson, 2010). These factors may influence lexical choice too, but prosody refers to how these phonetic or phonological features directly affect the speech signal since prosody is "an integral part of the phonological representation of speech" (Arvaniti & Fletcher, 2020: 79). Prosodic boundaries always coincide with syntactic boundaries, although not necessarily the other way around (Nespor & Vogel, 1986). For example, the sentence "the little boy who lived down the lane" is grouped into three prosodic units: *the little boy* (noun phrase), *who lived* (verb phrase), *down the lane* (adverbial phrase). Since perceiving the prosodic boundaries of an utterance through features such as phrase-final lengthening, pauses, stress or pitch draws notice to these boundaries in the acoustic speech stream, the contingent syntactic boundaries are also rendered more salient. This may help listeners to parse the continuous speech stream by dividing it up into salient chunks that contain meaningful syntactic units, such as words and phrases.

Prosodic parsing is an essential element of production and perception in all languages (Gussenhoven & Chen, 2020): speech cannot be produced without patterns of duration, pitch and intensity, and prosodic structure is used in planning utterances for production (see Shattuck-Hufnagel, 2020, for a detailed account). Prosody can also indicate breathing patterns, with breaths often taken at clause and sentence boundaries during spontaneous speech (Winkworth, Davis, Adams & Ellis, 1995), and coordinated with speech planning processes (Székely, Henter, Beskow & Gustafson, 2020). Discourse markers (e.g., turn taking) and emotional states are also marked by prosody (see Hirschberg et al., 2020, and Swerts & Krahmer, 2020).

In terms of form, in addition to tone (a segmental language feature, Gussenhoven & Chen, 2020), prosody consists of suprasegmental features of speech that are not confined to individual segments (i.e., consonants and vowels) but operate over larger units such as syllables, words and phrases (Lehiste, 1970). Suprasegmental features convey meaning,

affect, emphasis and the structural organisation of speech in communication. Key elements include **intonation**, the rise and fall of pitch across a phrase or sentence (e.g., the rising intonation that indicates a question); **stress** placed on certain syllables or words that may indicate content (nouns/verbs) or function words; **rhythm**, which is the timing and flow of speech and how syllables are grouped or spaced (e.g., the stress-timed rhythm of English or the syllable-timed rhythm of French); **pitch**, measured by the frequency of acoustic properties of speech, which can be higher or lower and contributes to tone and intonation; **speech rate**, which is the speed at which someone speaks; and **pauses**, which can be used strategically to convey meaning (e.g., *Let's eat, Grandma!* contains a meaningful pause before the noun[8]) and at clause boundaries.

A concern with listing prosodic features in this way is that the speech signal is not easily divided into its segmental and prosodic components to be analysed separately, as a list such as this may suggest. Information in the speech signal is encoded through fundamental frequency, duration, and intensity for both segmental and suprasegmental components (Wagner & Watson, 2010). There is an intrinsic coherence between prosodic and segmental aspects of speech (Wagner & Watson, 2010; Xu & Liu, 2012); it is hard to separate affective and linguistic aspects of prosody (Ní Chasaide & Gobl, 2004); and the perception of emotional speech involves the integration of prosody and semantics (Ben-David et al., 2016). We should thus avoid being misled by listing features into thinking one element of prosody is perceived in isolation from other aspects of the speech signal. An utterance is an holistic production process and is perceived holistically, even if individual prosodic features contribute to distinct parts of the speech act as outlined here (Wagner & Watson, 2010).

[8] Much to Grandma's relief!

2.3.1.3.2 Prosody, meter and music

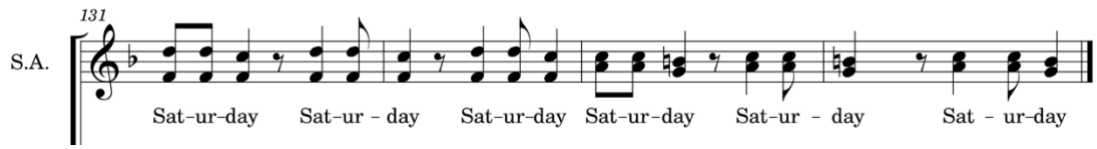
This section touches upon the key principles of meter and how it is different to prosody. It is helpful to understand the basic principles of meter to see how text-setting in songs may (or may not) represent natural speech patterns, which is relevant to the FL curriculum's assertion that songs are a way of exposing YLLs to 'the patterns and sounds of language' (DfE, 2013a). As outlined in 2.3.1.3.1, prosodic patterns are inherent organisational features of the speech signal that comprise prominence-defining acoustic elements such as stress, intonation and pitch contours. Metrics are superimposed onto this existing linguistic rhythmic organisation, bringing another layer of deliberate and formal stylistic constraints to the inherent prosodic patterns of language (Kiparsky, 2020). Meter can be thought of as designing metrical patterns, which are constraints on the quantity and stress patterns of syllables in a line of verse, and are themselves constrained by the grammar of language. Rhymes are often used to enhance metrical forms too (Kiparsky, 2020). Ultimately, meter exploits (and prosody arises from) our mammalian bias towards perceiving binary alternating constituents of rhythm in a sequence of beats, even when the sequence itself is a series of identical units (e.g., we perceive 'tick-tock' pairs in sequences of identical continuous 'tick-tick-tick' beats on a clock; Hayes, 1995; Kiparsky, 2020). The more prominent and less prominent beats alternate in either strong-weak (trochaic) or weak-strong (iambic) feet, through words or dipods, phrases or lines, and so on up the prosodic or metric hierarchy to the final level of utterance or poem respectively (Liberman & Prince, 1977; Hayes, 1995; Kiparsky, 2020). Mismatches between the structure of feet and phrasing are avoided in verse (Kiparsky, 1977; 2020) and speech (Hayes, 1995).

Human perception of regular linguistic rhythm is biased towards perceiving iambic or trochaic pairs, as predicted by the Iambic-Trochaic Law (Hayes, 1995), which was first developed as a theory of rhythmic grouping in music in psychological experiments (Bolton, 1894; Woodrow, 1951). To test this theory empirically, listeners are presented with two

series of artificial sounds with a regular rhythm and contrastive prominence based on either 1) alternate sounds being louder or 2) alternate sounds being longer. When asked to judge how to group the sounds into pairs, the usual preference regarding intensity contrast is for the louder (prominent) element to be first; for durational contrast, listeners put the longer (prominent) sound second. Thus, preferred perception patterns follow the law of well-formed rhythmic structure known as Iambic-Trochaic Law (ITL; Hayes, 1995).

Musicians adhere to ITL when transcribing verse rhythm in musical notation (Hayes, 1995). Indeed, composing a song with lyrics and melody involves aligning the prominent beats of the three independent rhythmic structures that arise from the inherent language prosody, the chosen poetic meter, and the musical rhythm (Kiparsky, 2020). Prominent phonological phrases and words are aligned in a way that matches the metrical constraints (respecting the metric feet, dipods and lines, etc.), and also the strong musical beats. English stress patterns are salient at the word level, and matched preferentially with strong musical beats. Imagine humming the outro of *Saturday night's alright* by Elton John[9] (1973) to get the gist – it would sound misaligned if the strongest musical accent fell anywhere but on the first syllable of the repeated word *Sat*urday, the strongest stressed syllable. Figure 2.1 shows how the stressed syllable does not always align with the first beat of the bar, but there is a syncopated rhythm, which elevates the prominence and salience of the stressed syllable because it is always slightly off-beat. The song is about unpredictable and even violent disruptive behaviour of young people, hence the syncopated rhythm of the music highlights the message in the lyrics.

Figure 2.1 Musical score for Elton John's Saturday Night's Alright



[9] Spotify link: https://open.spotify.com/track/12yHvSYFXI7PGzNecUvIDu?si=5728e6db14ce4a1b

For French, which is a syllable-timed language (more about this in section 2.3), beats are traditionally paired with syllables and stress is matched with strong beats only at the end of lines (Dell & Halle, 2009). Now imagine humming a few bars of the celebrated waltz *La Bohème* by Charles Aznavour[10] (1965):

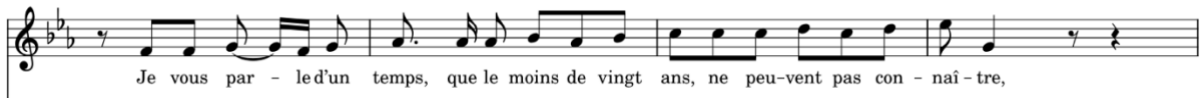
Je vous parle d'un temps

Que le moins de vingt ans

Ne peuvent pas connaître

The most prominently stressed final syllables align with the end of each line apart from in *connaître* which ends in an unstressed vowel (/kɔ.nɛtʁ/) in spoken French, but to respect the meter of the verse is pronounced with a schwa (/kɔ.nɛ.tʁə/) in the song. This final-stress structure aligns with the beginning of the bars in the musical score (Figure 2.2). There is a secondary stress on the third syllable in each line, creating a predictable three-four-time waltz rhythm in the verses that is then interrupted by shorter repeated phrases (*La Bohème, La Bohème* | *ça voulait dire* | *on est heureux*) in the chorus, where each syllable aligns with notes that are twice the duration of those in the verses. The disruption of the fast-paced regular verse structure and change to a slower pace in the chorus adds dramatic contrastive effect to the lyrics, which are about nostalgia for the romance of youth, when viewed from the perspective of an older man.

Figure 2.2 Musical score for Charles Aznavour's La Bohème



[10] Spotify link: https://open.spotify.com/track/1WvvmEowf7hiz5EnyAwtTj?si=191214f89a7a486at

Both of these examples demonstrate how English and French music respects the ITL and each of the three rhythmic systems (prosody, meter, music) are "mutually optimized" (Kiparsky, 2020: 673) to retain the most salient linguistic features. Taking a cross-linguistic perspective on the relationship between linguistic, metrical and musical prosody provides an insight into text-setting for songs and how the "patterns and sounds of language" (DfE, 2013a) can remain faithful to natural spoken language, or augmented stylistically to fit the meter with artistic effect, which may render them more prominent and salient.

Having had a brief excursion to look at the nature of prosody, and how it is put in service of art through meter and music, the following sections return to the prosodic bootstrapping hypothesis and how prosody helps infants learn their first language(s).

2.3.1.3.3 Bootstrapping accounts of L1 acquisition

The prosodic bootstrapping hypothesis is one of several 'bootstrapping' accounts put forward in psycholinguistic theory to address the 'logical problem' of how infants learn their L1 (see Höhle, 2009, for an historic overview). Sometimes called the projection problem, or the poverty of the stimulus, this logical problem refers to an apparent gap between the imperfect linguistic input infants receive and the complex linguistic knowledge they attain, with nativist accounts of L1 acquisition hypothesising that the solution must be pre-existing linguistic knowledge (Baker, 1979; Bley-Vroman, 1989; Chomsky, 1965, 1980; Hyams, 1988). This logical problem has given rise to several decades of research into the 'principles and parameters' account of L1 acquisition, whereby bootstrapping mechanisms (e.g., syntactic bootstrapping, Grimshaw, 1981; Pinker, 1984, and semantic bootstrapping, Gleitman, 1990) link the language an infant is exposed to with universal grammar, their innate linguistic knowledge.

In the 1980s, nativist-leaning theorists began investigating the possibility that acoustic information in the speech signal could provide the 'missing link' between the input and the child's acquisition of syntax (Gleitman & Wanner, 1982; Gleitman et al., 1987;

Peters, 1983, 1985), with the idea being that prosodic markers in the speech signal would help children discern syntactic units. As Jusczyk (1997) notes, early prosodic bootstrapping accounts were complementary to nativist accounts, with acoustic cues helping to bracket the input (Morgan, 1986, 1990). The seminal volume *Signal to Syntax* (Morgan & Demuth, 1996) marks a turning point towards a more input-driven account of prosody's role in L1 acquisition. Rather than assuming innate linguistic knowledge exists to fill the gap posed by the logical problem, input-driven or usage-based accounts of L1 acquisition hypothesise that domain-general cognitive processes drive infants' L1 development process through inducing knowledge of linguistic structures and their communicative functions, helping them break into the specific language(s) they are exposed to through input in social communicative situations (Saffran, Aslin & Newport, 1996; Tomasello, 1992, 2003). In usage-based accounts, linguistic input is key to children's iterative L1 development (Tomasello, 2003). Prosodic bootstrapping provides an explanation of how infants employ their sophisticated speech perception abilities (Eimas, Siqueland, Jusczyk & Vigorito, 1971; Kuhl, 2004) and use different aspects of linguistic prosody in the input to scaffold their learning of lexis and morphosyntax (Gervain, Christophe & Mazuka, 2020; Jusczyk, 1997). Rhythm (Vihman, Davis & DePaolis, 1995) and stress (Cutler & Foss, 1977) arguably provide a framework for acquisition because (along with vowel length, and other phonological processes) they create predictable temporal acoustic cues (Allen & Hawkins, 1980) that infants can exploit using domain- and species-general auditory processes and learning mechanisms (Hauser et al., 2002; Hauser et al., 2001; Newport et al., 2004; Ramus et al., 2000; Saffran et al., 1999).

Jusczyk (1997) outlines three conditions that must be met to associate prosodic bootstrapping with L1 acquisition. Firstly, there is evidence that acoustic correlates of syntactic markers are present in speech (Cruttenden, 1986; Gussenhoven & Chen, 2020; Pierrehumbert, 1980; Selkirk, 1984; see also section 2.3.1.3.1). Secondly, infants can detect these prosodic correlates in speech and, thirdly, they (at least partially) rely on prosodic cues

for organising the input. The following section presents evidence of infants' perception of and reliance on prosodic cues for bootstrapping their first language(s).

2.3.1.3.4 How infants perceive and exploit prosodic cues in the speech signal

Prosody is one, but by no means the only, source of information from the acoustic signal that infants rely upon for perceiving, differentiating and acquiring the languages in their environment. This thesis focuses on the verbal modality but the acquisition of sign languages follows a similar multi-faceted process (Mayberry & Squires, 2006; Rowland, 2014). There are a "constellation of cues from the speech signal" (Jusczyk, 1997: 141) providing information about how utterances are organised and produced that help infants begin the lengthy process of L1 acquisition. Multiple overlapping processes involving acoustic stress patterns (Abboub, Nazzi & Gervain, 2016; Byers-Heinlein, Burns & Werker, 2010; Thiessen & Saffran, 2003, 2007), transitional probabilities of segments in natural (Jusczyk, Charles-Luce & Luce, 1994; Maye, Werker & Gerken, 2002) and artificial languages (Saffran et al., 1996), position of function words and word order (Benavides-Varela & Gervain, 2017; Bernard & Gervain, 2012; Shi, Werker & Morgan, 1999), pauses (Kemler Nelson et al., 1989), and clause boundaries (Jusczyk, 1989; Jusczyk et al., 1993; Mandel, Jusczyk & Masuka, 1992) are just a few of the 'clues' in the input that infants perceive and use for bootstrapping their way into language. This section explores how prosody contributes to L1 acquisition as part of this dynamic, multifactorial, longitudinal and iterative process that begins before birth (Jusczyk, 1997).

Foetuses first encounter the speech signal in the womb, from around 20 weeks of pregnancy once their hearing capacity is operational (Eggermont & Moore, 2012; Pujol, Levigne-Rebillard & Uziel, 1991). *In utero*, the speech signal is predominantly prosodic: the melody and rhythm of the signal is preserved but the sound waves traverse maternal tissues, producing a low-pass filtering effect that removes phonetic cues, particularly suppressing consonants (Abrams & Gerhardt, 2000). The vocalic cues, however, are maintained and

signal prosody through their pitch, intonation and duration. Hence prosody represents children's first encounter with linguistic input in utero (Gervain, Christophe, & Mazuka, 2020).

There is evidence that foetuses attune to complex auditory streams during the third trimester, and learn more than just a general preference for their native language (Mehler et al., 1988; Moon, Cooper & Fifer, 1993) or languages (e.g, English and Tagalog, Byers-Heinlein et al., 2010), their mother's voice (DeCasper & Fifer, 1980), familiar melodies (Granier-Deferre, Bassereau, Ribeiro, Jacquet & DeCasper, 2011), or stories that are familiar from being heard *in utero* (Kisilevsky et al., 2009). They also gain language-specific and detailed knowledge that is consistent with adult-like prosodic grouping preferences predicted by ITL (Hayes, 1995). Their prenatal prosodic experience biases them towards perceiving and organising sound sequences according to the acoustic patterns of the languages they experience prenatally (Abboub, Nazzi & Gervain, 2016). Newborns discriminate between familiar and unfamiliar vowel sounds in English and Swedish (Moon, Lagercrantz & Kuhl, 2013), and also vowel and pitch changes that differ from those experienced before birth (Partanen et al., 2013). Vocalic cues are important for signalling prosody, which in turn signals phrasal structure (Nespor & Vogel, 1986). Thus, newborns' ability to perceive and discriminate vowel sounds may help them take their initial steps towards learning syntax and lexis. As well as perceptive biases, newborn productions (i.e., cries) are consistent with the intonational and phonological phrases of the languages they hear before birth. There is some evidence (see section 2.3.5.1 for discussion) that German infants produce falling cry contours with an initial pitch peak, and French infants produce rising cry contours with a final pitch peak, both consistent with the prosodic patterns of the language (either German or French, see summary in Ommen et al., 2020) they heard *in utero* (Mampe, Friederici, Christope & Wermke, 2009).

Newborns can also use their perceptive skills to discriminate between rhythmically different languages[11]. Nazzi, Bertoncini and Mehler (1998a) found that French newborns discriminate between stress-timed English and mora-timed Japanese, or between English/Dutch sentences and syllable-timed Spanish/Italian sentences, but do not distinguish between utterances from languages within the same rhythmic class (e.g., English from Dutch, or Spanish from Italian). For rhythmically similar languages, such as Spanish and Catalan, infants appear to differentiate only from 3.5–4 months old, and to rely on phonotactic regularities and phonemes rather than prosodic cues (Bosch & Sebastián-Gallés, 1997). Extracting linguistic information using prosodic cues thus appears to be a foundational process that begins before birth and is used in synergy with other input cues like phonotactics and distributional analysis (Jusczyk et al., 1994; Maye et al., 2002; Saffran et al., 1996) to orient infants to their linguistic environments, where they may be exposed to multiple languages. They can then begin discriminating and categorising the acoustic cues that they hear.

Building on these foundations, infants' keen prosodic perceptive skills help them develop key areas of knowledge about language that may assist in L1 acquisition: knowledge about prosodic boundaries in utterances and clauses specific to the languages that they are exposed to; knowledge about stress or prominence patterns within and across languages; and knowledge about function words and content words. These three areas are mutually assistive

[11] There is some debate about the validity of the classic terms 'mora-timed', 'stress-timed' or 'syllable-timed' to categorise languages based on the equal rhythmic timing (isochrony) of their privileged phonological units (mora, foot or syllable, respectively) in speech production (Pike, 1945; Ladefoged, 1975) because there is inconclusive acoustic evidence that such unit-based timing differences exist (see Cutler, 1994, for a review, and Post & Payne, 2018, for an alternative, multi-factorial rhythm account of prosodic bootstrapping that does not rely on these three classes). Yet infant perception studies indicate that newborns do discriminate between languages in some way that aligns with these three categories. Ramus et al. (1999: 270) argue that rather than perception being based on phonological unit (mora, foot, syllable) isochrony, infants perceive "speech as a succession of vowels of variable durations and intensities, alternating with periods of unanalysed noise (i.e., consonants)." However, since the terms continue to be used and provide a useful shorthand, I shall use them also for the sake of simplicity.

in segmenting the speech stream to help babies develop understanding of lexis, morphosyntax, and word order, as explored in the following overview of evidence.

Knowledge about prosodic boundaries is helpful because prosodic contours frequently (but not wholly consistently) map on to meaningful linguistic units such as utterances or phrases (Cruttenden, 1986; Pierrehumbert, 1980; Selkirk, 1984). Newborn perception of prosodic contours enables them to demarcate meaningful units, such as words and phrases, in the continuous speech stream, even in languages they have not been exposed to during gestation. For example, French newborns (mean age, 68h) discriminate lists of two-syllable Japanese words that differ only in pitch (high-low, low-high) contrasts (Nazzi, Floccia & Bertoncini, 1998b). At two months, infants can detect word order changes provided they are embedded within a coherent prosodic structure (such as a sentence) not across a prosodic boundary (Mandel, Kemler Nelson & Jusczyk, 1996). Infants grow more familiar with the prosodic patterns that are specific to their native languages as they gain more linguistic experience. Younger infants appear to demonstrate a language-general sensitivity to prosodic marking of clausal units in speech utterances, which may become more language-specific as they tune into the prosody of the languages in their environment. For example, Jusczyk (1989) found that American 4.5-month-olds show a preference for hearing utterances where artificial one-second pauses are inserted at clause boundaries (coincident versions) rather than in the middle of clauses (non-coincident versions) in English (their native language) as well as in Polish, an unfamiliar language. But by six months old, the American infants no longer show any preference for the coincident versions compared to non-coincident versions of Polish samples, even when the samples are low-pass filtered to remove phonetic information and focus on the prosody. These findings indicate that Polish prosody marks it out as nonnative for the American infants at six months, whereas at 4.5 months they perceive prosodic marking of clausal units regardless of the input language's familiarity.

However, two similar studies with American 4.5-month-olds found no evidence that the infants listened significantly longer to coincident versions of Japanese utterances than non-coincident versions (Jusczyk et al., 1993; Mandel et al., 1992), which makes it less clearcut to state that 4.5-month-olds possess language-general perception of prosodic markers to clause boundaries. This could be because detection of prosodic cues such as phrase boundary markers is language-specific, or because languages differ in how they mark prosodic prominence through vocalic acoustic cues (Nespor & Vogel, 1984). For example, in English and French phrases, vowel duration marks the contrast between prominent (e.g., nouns and verbs) and non-prominent elements (e.g., function words) (Nespor & Vogel, 1984; Selkirk, 1984). In the phrase *to Rome*, the /o/ is short in *to* and long in *Rome* (and in French too, *à Rome*, there is a similarly short /a/ then long /o/ vowel). Thus the noun, *Rome*, is given prominence over the function word, *to* or *à* (Abboub et al., 2016). There is a pattern of phrase-initial function word and phrase-final content words in English and French, producing an iambic prosodic pattern of short-long (or weak-strong) pairs since function words are subject to reduced stress, duration and segmental complexity than content words (Hayes, 1995; Vaissière & Michaud, 2006). Japanese, on the other hand, marks prominence through pitch or intensity contrasts, as in ^*Tokyo kara* ('Tokyo to'). Here, the phrase-initial sound is louder or higher, making it more prominent, and sequences create a trochaic prosodic pattern of loud-soft (or strong-weak) pairs (Hayes, 1995).

Newborns are sensitive to prosodic contours at the utterance level, discriminating stimuli containing well-formed utterances from those containing reversed (thus deviant) prosody at just 1–3 days old (Martinez-Alvarez, Benavides-Varela, Lapillonne & Gervain, 2023), as well as at the word level (Nazzi et al., 1998a). At four months old, in preferential head-turn experiments, infants show a preference for passages with artificial pauses inserted at clause boundaries over passages with pauses inserted within clauses (Jusczyk, Hohne & Mandel, 1995). At nine months old, infants show sensitivity to smaller prosodic units

(Gerken, Jusczyk & Mandel, 1994) and at 13 months, they can use prosodic boundaries to constrain lexical access. For example, once trained to recognise the word *paper*, 13-month-olds reject sentences where the two syllables span a prosodic boundary, as in *[the man with the highest **pay**] [**per**forms the most]* (Gout, Christophe and Morgan, 2004). Similarly, French 16-month-olds show an ability to exploit phonological phrase boundaries to constrain lexical access (Millotte, Morgan, Margules, Bernal, Dutat & Christophe, 2011). Preschoolers can disambiguate phrases (e.g., *the baby flies* or *la petite ferme* [noun phrases] from *the baby / flies* or *la petite / ferme* [verbs]) based on prosodic boundaries in English and French (de Carvalho, 2017).

Prosody is also a useful cue for learning word order, working together with other cues such as word frequency. Languages vary in their word order conventions, namely where function words are positioned within phrases, and the relative order of verbs and objects. Function words are a closed class of items (including articles, pronouns, etc.), which indicate morphosyntactic sentence structure (Fukui, 2006). They are more frequent than content words, an open class of items carrying lexical meaning, and tend to appear in utterance-initial or utterance-final positions, and are perceptually salient to infants (Aslin, Woodward, LaMendola & Bever, 1996). The position of function words corresponds to word order. French, Italian and English are VO or functor-initial languages [*de Rome/di Roma/from Rome*]. Japanese is an OV or functor-final language [*Tokyo kara* – 'Tokyo from']. Japanese and Italian have opposite word orders (OV for Japanese and VO for Italian), and monolingual infants exposed to these languages show a preference for their familiar word order in artificial grammar learning tasks (Gervain et al., 2008). Some languages have both object-verb and verb-object word order (e.g., German and Dutch). Thus bilingual (e.g., Japanese/Italian) infants could be exposed to two word orders and monolingual German/Dutch infants will hear both OV and VO orders intra-lingually. Infants therefore cannot only rely on function word frequency and position to parse the speech signal. They

require another cue from the input in order to learn word order. Prosodic phrasing provides one such cue.

Prosodic prominence location (at the head or tail of intonational phrases) and acoustic realisation (through stress, tone and duration) indicate function and content words, providing infants with clues to underlying linguistic structure (Morgan & Demuth, 1996). In French and English, function words appear before nouns and verbs (since these are head-initial languages) and infants are sensitive to function words from early on (Hallé, Durand & Boysson-Bardies, 2008; Shi et al., 1999, 2006). Syntactic phrases often begin with a function word and end with a content word in English and French, and the edges are acoustically rendered with prosodic phrasing, where the function word is minimally prominent and the content word is prominent (Hayes, 1995; Liberman & Prince, 1977). It seems that edge words are thus more salient and easier to segment from continuous speech than words the middle of utterances (Cutler, 1994; Endress & Mehler, 2009; Johnson, Seidl & Tyler, 2014). Toddlers of 14–24 months old use their knowledge of function words to constrain lexical access, expecting a noun after a determiner for instance in French (Cauvet et al., 2014) and English (Kedar, Casasola & Lust, 2006). The combination of phrasal prosody and function words constrains syntactic analysis, which in turn constrains the possible meanings of novel lexical items (de Carvalho, 2017; de Carvalho, He, Lidz & Christophe, 2019).

Converging evidence suggests that by using prosodic cues and function words, in combination with emerging knowledge of a handful of lexical items (the 'semantic seed'; Gutman, Dautriche, Crabbé & Christophe, 2015), infants can infer syntactic categories (i.e., noun phrases and verbal nuclei), and use that knowledge to further enrich their vocabulary, which then further expands their syntactic knowledge in this mutually supportive, iterative process of language acquisition (Gussenhoven & Chen, 2020). Thus, whilst L1 acquisition is highly complex, according to a prosodic bootstrapping account, the input contains the necessary acoustic cues required to begin the process of segmenting and categorising the

speech stream (Gleitman & Wanner, 1982; Morgan & Demuth, 1996; see section 2.3.4.2 for more on infant-directed speech). All this happens in combination with social and emotional cues in a caregiving context or dyad (Levine, Hirsh-Pasek, & Golinkoff, 2020; Tomasello, 2003) and is powered by learning mechanisms (e.g., distributional learning) that are general rather than species- or domain-specific (Hauser et al., 2002; Hauser et al., 2001; Newport et al., 2004; Ramus et al., 2000; Saffran et al., 1999). There are still open questions about which prosodic cues are the most useful at different stages of development and for which languages, how bilingual infants employ prosodic bootstrapping processes across typologically different languages, and how prosodic cues interact with statistical learning and phonotactics (Gussenhoven & Chen, 2020; Prieto & Esteve-Gibert, 2018). However, it seems likely that the prosodic bootstrapping mechanism is a "powerful heuristic" (Gervain et al., 2020: 573) in L1 acquisition. The following section considers how this account of prosodic bootstrapping may apply to L2 learners in classroom FL contexts, as opposed to infants in naturalistic L1 acquisition contexts.

2.3.1.3.5 Does prosodic bootstrapping apply to young L2 learners?

The evidence reviewed in section 2.3.1.3.4 indicates that infants are able to perceive prosodic cues such as intonation, stress and duration in the speech signal, and use these to piece together the overall structure of utterances, discriminate between content and function words, and begin to bootstrap the lexicon and syntax of their L1. The perceptual and learning mechanisms involved are present from birth (or indeed before) and there is evidence of the continued use of prosodic cues to parse utterances until five years old (de Carvalho et al., 2016). For speech production, the process is more drawn out due to physical and motor constraints, with French-speaking children acquiring adult-like speech prosody around age two, but children still grappling with producing adult-like prosody at age six in stress-timed languages such as English (see Post & Payne, 2018, for a review of relevant production evidence). Prosody is also vital for pragmatic and socio-emotional processing of speech

input, with discourse-level skills developing over the longer term from age 3 to 11 years (see Chen, Estève-Gibert, Prieto & Redford, 2020, for a review). The L1 prosodic system is thus well-established and actively involved in multiple L1 acquisition and production processes throughout childhood, and into adulthood prosody continues to influence speech parsing (Streeter, 1978) and disambiguation (Price, Ostendorf, Shattuck-Hufnagel & Fong, 1991).

When it comes to learning second languages, there is evidence that prosody is used for speech segmentation (Pilon, 1981) and identification of phrasal structure (Wakefield, Doughtie & Yom, 1974), but that the existing L1 prosodic system may influence adults' L2 word perception and recognition (Jongman & Tremblay, 2020), and their production (Goad & White, 2019; Trouvain & Braun, 2020). L2 learners with an L1 where stress is lexically contrastive are more likely to use their L1 perception of stress when learning words in an L2 than learners whose L1 does not have lexically contrastive stress, for example French learners of Spanish (Dupoux, Pallier, Sebastian & Mehler, 1997; Dupoux, Peperkamp & Sebastián-Gallés, 2001) or English (Tremblay, 2008; Tremblay, Broersma, & Coughlin, 2018). This has led to some accounts of 'stress deafness' for French L1 speakers (e.g., Dupoux et al., 1997, 2001, 2008, 2010) for example, when parsing words in an L2 that does mark lexical contrast with stress because the L1 parameter for stress perception has been set and influences L2 perception. However, the picture may be more complex and nuanced than accounts of 'stress deafness' suggest, with more factors than the broad categories of language rhythm typology at play.

There is evidence that L2 learners (much like infants learning their L1) require an integrated and multi-faceted command of multiple areas of grammar, syntax, and lexical domains or interfaces between these domains for successful L2 acquisition (Sorace, 2011). Focusing in particular on the level of the prosodic word and phonological phrase, thus on acquisition of morphosyntax, The Prosodic Transfer Hypothesis (PTH; Goad & White, 2004) posits that prosodic structure transferred from L1 grammar constrains L2 perception,

comprehension and morphology production (i.e., inflections and function words). PTH has undergone several revisions since the initial strong 'full transfer no access' position (Goad et al., 2004), in which L2 learners' interlanguage grammar demonstrates they are permanently 'stuck' with L1 representations. In the strong version of PTH, L2 learners are unable to represent or produce L2 tense morphology, (e.g., in the case of Mandarin speakers learning English, Hawkins & Liszka, 2003) or function words (e.g., Turkish speakers learning English, where stress patterns for functors differ and L1 representations override L2 patterns resulting in omission, in determiners being used instead of articles, or non-target-like pauses in phrases; Goad & White, 2009). A more moderate 'full transfer partial access' PTH account followed, whereby L2 learners could build representations of L2 verb inflections (Goad & White, 2006) or plural morphology and articles (Goad & White, 2004) by minimally adapting L1 representations (see Goad & White, 2019 for a summary). Finally, in the 'full transfer full access' conception of PTH, learners' L1 prosodic transfer constrains some stages of L2 acquisition but it is not necessarily permanent and can be unlearned in some cases, potentially depending upon individual differences such as proficiency level. For example, Turkish does not support free clitics as English does in the construction [article] + [adjective] + [noun] because the Turkish article (*bir*) would be stressed before the adjective, whereas in English the article (*a*/*the*) remains unstressed, resulting in non-target-like prosody when rendered using the Turkish prosodic phrasing (Goad & White, 2009). However, on re-examing their data, Goad and White (2009) found that two individuals did produce target-like performance of the English articles, which raised the question of whether standardised proficiency tests provide accurate assessment of learners' phonological ability, which would influence their performance, or focus on lexical and syntactic knowledge that may not wholly represent their ability to produce target-like article stress (Goad & White, 2019). There thus appear to be some exceptions to the PTH that may render it less explanatory of how prosody is transferred from L1 to L2.

Perhaps this lack of consistency in PTH experimental data is because PTH is founded on a nativist account of language acquisition, and thus assumes learners build new representations of L2 grammatical knowledge in their interlanguage that may compete with established L1 representations, resulting in non-target-like productions (Goad & White, 2019). However, if one takes a usage-based foundation for language acquisition as the starting point, it could be argued that L2 learners have simply not experienced enough input to build target representations and that focusing on input where L2 prosody is more salient or taught explicitly would help learners perceive and produce target language prosody.

Indeed, there is evidence that adults can use non-native prosodic prominence cues within prosodic phrases to learn novel lexical items and word order in an artificial language, but only after a certain amount of exposure (Saksida, Flo, Guedes, Nespor & Garay, 2021). Furthermore, teaching L2 prosody and suprasegmental features explicitly may improve fluency and comprehensibility of L2 learners' speech (Gordon & Darcy, 2016). And, finally, there is a handful of evidence that prosodic bootstrapping processes extend to assisting YLLs in classrooms with L2 word-order acquisition and elicited imitation tasks (Campfield & Murphy, 2013; 2014), although both of these studies were conducted with Polish primary school EFL learners, hence it remains to be seen whether the findings transfer to other contexts. It could be the case that presenting YLLs with L2 input through singing or verse 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. However, there does not appear to be much (if any) research building on the prosodic bootstrapping hypothesis with YLLs in FL contexts in the UK.

2.3.3 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, they provide

unstable foundations for experimental work to build upon because they lack robust empirical support to link them with linguistic outcomes. 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, an area that has received little attention to date in UK FL contexts.

2.3.4 Evidence from classroom experiments

This section reviews evidence from experiments with learners who are engaged in learning their L1 or their L2 in classroom contexts, where songs were investigated for their influence on the learners' linguistic development.

2.3.4.1 From L1 classrooms

Several[12] empirical studies investigate using songs for L1 literacy and vocabulary acquisition in classroom settings. Joyce (2011) investigated whether singing along with two different stories whilst they were being read benefitted children's productive vocabulary learning. 71 children aged 5–7 (68 English native speakers and three ESL learners) were allocated in intact classes to story-with-song or story-only conditions, with a crossover design. Target vocabulary was presented as pictures at pretest (elicited oral recall task of target items), during the 30-minute interventions (two sessions of reading/singing the stories over two days), and again at posttest. Participants all learned new vocabulary, but they learned no more through the singing than the read-only condition. However, Joyce did not conduct baseline cognitive tests and the non-randomised allocation of intact classes to conditions makes the sample size $n = 4$, one class per condition, not $n = 71$ individual cases. Given the methodological shortcomings of this study, it is difficult to assess the applicability of its findings to other contexts. In an interesting insight into teachers' opinions about using songs,

[12] This section has been adapted from Hamilton & Murphy (2023). In that paper, the first author conceived and designed the study, collected the data, conducted the analysis, and wrote and edited the paper. The second author supported the conceptualisation and edited the final paper.

one teacher reported that her students were more engaged in the song condition. However, this was contradicted by the researcher's observations, suggesting that teachers may believe songs are more engaging for students than other methods even if this is not demonstrated empirically.

Two further studies investigate the impact of structured music programs (including singing activities) on 3–4-year-olds' receptive and expressive language development (Crosswhite, 1996) and 5–6-year-olds' phonological awareness skills (Lehman, 2019). Crosswhite (1996) found no detectable effects of music on language skills, measured using a receptive vocabulary test (Dunn & Dunn, 1981) and language sample analysis of mean length of utterance. Lehman (2019) found a positive effect of song input on rhyming production and segmentation (sentences, syllables and phonemes) (PAT2; Robertson & Salter, 2007). However, their quasi-experimental designs (with small samples) and lack of baseline cognitive tests make these results unreliable. Crosswhite's null result could be explained by failing to control for the comparison group's musical exposure, since music is a feature of all preschools in some form (Hamilton & Murphy, 2023; Siraj-Blatchford, 1999).

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. Overall, scant reliable classroom evidence is available to support claims about songs' influence on L1 literacy or linguistic measures, and the studies discussed here are not generally applicable to L2 YLL contexts because their participants are learning their L1 in naturalistic contexts, not an L2 in input-limited contexts. The next section reviews studies of L2 learners where singing was investigated.

2.3.4.2 From L2 classrooms

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 sparse 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 without an evaluation of the methodological quality of the included studies, little can be understood about the robustness of the findings. Overall, then, a transparent, systematic, and replicable approach to evaluating the state of the knowledge is needed.

2.3.5 Transdisciplinary evidence

In addition to Campfield and Murphy's (2013; 2014) theoretical contribution that prosodic bootstrapping 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. This section first gives an overview of research investigating melodic complexity and language-specific infant pre-speech and cry vocalisations. Then follows an overview of research looking at the prosodic content of infant-directed speech and similarities with singing, as well as the ubiquity of songs in child-directed input. Finally, research that links the auditory (thus physically-sensed) properties of prosodic input such as verse and songs with encoding patterns in the brain are explored. Whilst not directly related to L2 learners in formal educational contexts, these additional areas of research provide further motivation for this doctoral study.

2.3.5.1 Infant pre-speech and cry melodies

Whilst section 2.3.1.3 explored how prosody influences L1 acquisition with reference to the prosodic bootstrapping hypothesis, it did not address how infant pre-speech production

might be influenced by the prosody of their ambient (environmental) languages. One strand of research investigating infant cries proposes that prenatal and newborn experience of linguistic input is strong enough to shape pre-speech production (Gervain, 2018). In a series of studies, Wermke and colleagues posit that musical elements of infant pre-speech (in particular their cries, but also non-cry vocalisations [vocants; Martin, 1981]) are a necessary stage in language acquisition, rather than a by-product, and their increasing melodic complexity can be traced through crying, cooing and babbling stages of pre-speech (Wermke & Mende, 2009, 2016; Wermke, Robb & Schluter, 2021). A further strand of cry research investigates the link between ambient language prosody and infant cry melodies (Mampe et al., 2009; Wermke et al., 2016, 2017; Prochnow, Erlandsson, Hesse & Wermke, 2019). This section will present evidence from these two strands of pre-speech research and how it provides impetus for further investigation into the link between songs and language acquisition.

The first strand of research involves the Melody-Development Model (Mampe et al., 2009; Wermke & Mende, 2009, 2016; Wermke et al,. 2021), a complexity hypothesis where infants' early vocalisations iteratively develop from simple into more complex combinations of melody arcs during the phases of pre-speech and speech, relating infant cries to singing. Further support for increasing complexity in infant cry 'melodies' comes from research analysing visuals of frequency spectrograms of vocalisations produced at 3, 6 and 9 months (Kent & Murray, 1982) and a longitudinal qualitative study of speech quality in 2–6-month-old infants' non-cry vocalisations (Hsu, Fogel & Cooper, 2000). However, neither of these earlier studies looked at newborn cry data and Hsu et al., (2020) coded the melody arcs impressionistically by listening to them, rather than looking at spectograms, which may have been unreliable.

Wermke and Mende (2009) attempted to produce a more comprehensive report investigating the development of melodic arcs in cry vocalisations from birth to 12 weeks,

based on a series of earlier and ongoing studies (Mende et al., 1990; Wermke & Mende, 1994; Wermke et al., 1996; Wermke, 2002; Wermke & Friederici, 2004). Wermke et al., (2021) extended this investigation to infants aged six months. Wermke and colleagues operationalise a melody arc as a pitch (F0) contour *glissando* longer than 150ms, a smooth vocalisation that glides over a frequency interval of three semitones for crying and two semitones for other vocalisations (Wermke et al., 2021). Wermke and colleagues define complexity as an increase in how many melody arcs are produced within a cry sequence, over the first few months of life.

The infant vocalisations in Wermke et al. (2021) were taken from a database of baby sounds from typically developing infants at the University of Würzburg, Germany, with 67,629 vocalisations from 277 infants included (56,537 cry vocalisations from 227 infants and 11,092 non-cry vocalisations such as babbling and cooing from 50 infants, with one infant appearing in both audio datasets). Cry melodies were divided into simple (single melodic arc) and complex (multiple-arc) melody, and bubble plots produced to visualise the proportion of complex melody vocalisations recorded by age (measured in days). Further multi-level mixed effects logistic regression models for the simple/complex melody patterns (binary data) to account for the repeated vocalisation data measurements nested within children over time were conducted. The analyses reveal that infant cry and non-cry vocalisations become more prosodically complex from birth to 180 days in a curve: there is rapid complexity development (meaning a higher proportion of cries demonstrate a complex structure) over the first month, with 53% of cry vocalisations demonstrating complex melodic arcs at 30 days. The instances of complexity increased up to 4.5 months (140 days), after which there was observed to be a slight decrease – perhaps due to establishing new patterns of vocal development in consonant–vowel syllable sequences in babbling. Further research into the interaction between vocants/closants and melodic complexity are proposed to build on this statistical model demonstrating melody development in infant vocalisations.

The second strand of research investigating infant cries has looked for group similarities in fundamental frequency contour variation ('melody arc') according to the ambient languages present for French and German infants (Mampe et al., 2009) or for German and Cameroon (Nso) infants (Wermke, Teiser, et al., 2016). This research is grounded in studies that find newborns have a perceptual preference for their ambient native languages (e.g., Byers-Heinlein et al., 2010; Mehler et al., 1988; Moon et al., 1993), which raises the question of how early in development newborns can produce the sound systems of the languages familiar to them. To address this question, Mampe et al. (2009) analysed the crying patterns of 30 French and 30 German infants (aged from 2–5 days old). They collected 2500 cries, operationalised as vocal output produced in a single expiration, with 1254 cries included in the final between-groups analysis based on language group. In the French group, between 3 and 54 cries were analysed per newborn (mean count = 21), and in the German group between 10 and 38 cries (mean = 18). The wide individual count range was due to avoiding any elicitation or stimulation to provoke crying. All cries were produced spontaneously during routine mother-child caring interactions such as calming when fussy or changing nappies. The authors present data analyses showing a statistically significant between-groups difference in cry melody 'arc' that they claim reflects an early impact of the respective intonation patterns of their native languages. Newborn cries reached a maximum pitch (F0) at $M = 0.44$s ($SD = 0.15$s) for German, and $M = 0.58$s ($SD = 0.13$s) for French infants (Mann-Whitney test statistic not provided, $p < 0.0001$). The observed cry patterns are consistent with falling prosodic pitch contours across words or phrases observed in German utterances (Wiese, 1996) and rising prosodic pitch contours for French (Delattre, 1961; Welby, 2006). Mampe and colleagues attribute the group differences to the newborns' readiness to reproduce the prosodic patterns in their respective ambient languages.

These findings are consistent with a follow-up study by Wermke, Teiser et al. (2016) which compared 21 German infants and 21 infants from Cameroon who were from a

Lamnso-speaking Nso population (mean age = 4 days). Lamnso is a mono-syllabic eight-tone language and has multiple F0 patterns (Grebe & Grebe, 1975), in contrast to German which is multi-syllabic with predominantly falling F0 contours (Wiese, 2000). Wermke and colleagues tested the hypothesis that the F0 contours produced in the 1279 recorded cries would vary significantly differently between infants in the two language groups, with higher variability in the Lamnso group due to the complex tonal input they would hear. Similarly to Mampe et al. (2009), there was a wide range of cry count for individual infants due to deliberately avoiding eliciting cries. There were 26 mean cries recorded per German infant (range 5–42) and 22 (3–83) per Cameroon infant, with 1002 cries included in the final MANOVA (multiple analysis of variance) across F0 measures and cry duration. The authors report a significant multivariate effect of language group (Wilks' $\lambda = .65$, $F(4, 37) = 3.96$, $p < .01$, $\eta p^2 = .36$) and significant post-hoc analysis group differences on cry duration ($F(1, 40) = 6.50$, $p < .05$, $\eta p^2 = .14$) and three F0 measures: F0 range, $F(1, 40) = 5.38$, $p < .05$, $\eta p^2 = .12$, pitch sigma, $F(1, 40) = 8.72$, $p < .01$, $\eta p^2 = .18$, and F0 fluctuation, $F(1, 40) = 7.36$, $p < .05$, $\eta p^2 = .16$, but not mean F0 between groups, $F(1, 40) = .78$, $p > .05$, $\eta p^2 = .02$. Cameroon infants thus apparently produced longer cries with more pitch variation than German infants, but average pitch (F0) was similar between groups.

The findings from Wermke et al. (2016) are commensurate with the findings from Mampe et al. (2009) and further similar studies that found significant group differences using similar methods and analysis with larger samples comparing 6480 cries from 102 German and Mandarin infants (Wermke et al., 2017) and 6687 cries from German and Swedish infants (Prochnow, et al., 2019). Taken together, this series of similar studies with converging findings could indicate a promising additional avenue of prosody-related perception and production research that is potentially relevant to this doctoral investigation.

However, strong criticism of the statistical methods used (Gustafson, Sanborn, Lin & Green, 2017) urges caution in over-zealously interpreting the findings of Mampe et al.

(2009), Wermke et al. (2016, 2017) or the more recent Prochnow et al. (2019) studies as evidence of language-specific influences of prosody on newborns' production. Gustafson et al. (2017) critique the findings outlined above that newborn cries demonstrate significant group differences based on ambient language background. The critique is justified because the Mampe et al. (2009) and Wermke et al. (2016, 2017) statistical analyses treated each cry as an independent data point, where – given the well-documented within-infant similarities of cries, infants' oral "signature" to elicit care from their parents – a multilevel model is necessary to avoid violating the assumption of independence required to produce valid results from the statistics Mampe, Wermke and colleagues carried out. Indeed, neonate cries have been found to be robustly individual – mothers can recognise their own infants' cry (Formby, 1967; Green & Gustafson, 1983), as can fathers (Gustafsson, Levréro, Reby & Mathevon, 2013), and non-parent (nulliparous) adults can be trained to recognise individual babies' cries (Green & Gustafson, 1983). Gustafson et al. (2017) thus questioned the reliability of the findings from Mampe et al. (2009) and Wermke et al. (2016, 2017) because in these prior studies, cries from infants were treated as independent data points rather than as nested data, as would have been appropriate to produce valid results.

To illustrate their point using similar data gathering and analysis methods, Gustafson et al. (2017) carried out a two-part investigation of differences in neonate crying characteristics between American and Chinese infants using two models. In the first analysis, each *cry* (n = 497) was treated as an independent data point, and group differences were analysed with a t-test. The second analysis treated each individual *child* as the unit of analysis, with individual cries (n = 15 per child) treated as nested data within the child level. Their findings indicate a clear tendency to overstate group differences and declare false positive results when individual cries are treated as independent datapoints. The multilevel model, where individual variance is accounted for by ascribing the child as the unit of analysis, produces only one significant 'group' difference (SD of F0 between groups, with

English [non-tonal ambient language] infants showing greater tone variability than Mandarin [tonal ambient language] infants). Gustafson et al. (2017) attribute their finding to chance, especially since Wermke et al. (2017) found the opposite: that their Mandarin infants had significantly more tonal variability than German infants.

In spite of the flaws in statistical analyses that may invalidate their key findings, the series of studies from Wermke and colleagues into language group differences in neonate cry intonation continues to be a widely shared research headline by journalists in the *Daily Mail* (Derbyshire, 2009) and *New York Times* (Hardach, 2020), and on social media. One reel on Instagram by knoxstudy (2023) describing the studies has almost 80,000 likes and 900 comments, many of them about how intuitive this research finding is as it chimes with viewers' existing beliefs. One commenter also links it to the Mozart effect by saying "They do *[Newborns do have accents]*. That is why I played baby Mozart for them. 💜💜💜💜💜💜💜💜 " (ana_loves_this_world, 2023). There appears to be a synergy between the intuitive belief in the power of music and pre-natal linguistic development that appeals to popular cultural beliefs.

In the absence of a follow-up paper (to the best of my knowledge) for the language-group differences in cry melodies with a more robust statistical analysis of the data (as would appear to be possible, given the success of the multi-level model in Wermke et al., 2021), the combined findings of these two research strands on melody complexity development within languages (specifically German) and differences in cry melodies between languages (German and French, Mandarin, Nso, Swedish) provide tentative evidence that prosody influences very young children's pre-speech vocalisations. Perhaps this only provides contextual detail in this thesis, but pending further (more robust) analysis of language-specific cry vocalisations, it would be interesting to see how far into childhood these prosodic perception and production mechanisms extend, and whether they extend into

L2 acquisition in formal educational settings where children listen and repeat in an input-limited context remains an interesting and open question.

2.3.4.2 Infant directed speech and singing

Chomsky (1965: 32) famously referred to the input children receive as "restricted in scope" and "fairly degenerate in quality," also noting that "much of the actual speech observed consists of fragments and deviant expressions of a variety of sorts" (p.215). Researchers soon began accumulating robust evidence that Chomsky had mischaracterised children's linguistic input (Chouinard & Clark, 2003). Indeed, far from consisting of a systematically poor model of the linguistic system, infant-directed speech (IDS) is routinely modified linguistically and prosodically by caregivers to adapt and simplify language in ways that help children acquire it. This section explores how infant-directed communication (speech and singing) differs prosodically from adult-directed communication; how babies' preference for infant-directed styles of communication has been researched and the implications of methodological limitations; the potential functions of IDS; and how IDS could provide a foundation for learning FL through songs.

Infants, like adults, can perceive the prosodic marking of syntactic boundaries in the speech signal (see section 2.3.1.3.4), which are marked by changes in pitch, pauses, syllable duration and phrase-final lengthening (Nespor & Vogel, 1986; Price et al., 1991). In speech that is uttered to children, known as infant- or child-directed speech, these prosodic cues are exaggerated with shorter and simpler utterances, higher average pitch, wider pitch range, more prosodic repetition and longer content-word duration than in adult-directed speech (Snow & Ferguson, 1977; Fernald & Simon, 1984; Garnica, 1977; Stern, Spieker, Barnett & MacKain, 1983). These and other differences in prosody in IDS are well documented across multiple (but certainly not all) languages (Broesch & Bryant, 2015; Bryant, Liénard & Clark Barrett, 2012; Fernald & Simon, 1984; Fernald et al., 1989; Grieser & Kuhl, 1988; Räsänen, Altosaar & Laine, 2008), with evidence of both mothers and fathers (Papoušek, Papoušek &

Haekel, 1987) intuitively adjusting their prosodic register for their children. Papoušek and Hwang (1991) found that their Mandarin-speaking participants adjusted their pitch and intonation even when they were merely simulating speaking to infants or role-playing an FL teacher who wants their student to understand and reproduce what they say. It seems reasonably clear, then, that adults adjust their speech to accommodate the linguistic capacity of their interlocutors, including children and students, with IDS existing in some culturally-centered, sociolinguistically moderated form rather than a universally invariant speech style (Weinstein & Baldwin, 2024).

Regarding infant-directed song, a number of parallels can be drawn with the prosodic adjustments in IDS. Trehub, Unyk and Trainor (1993a) recorded American and Indian mothers singing one song to their infants and another without their infant present. The authors identify a distinctive style of singing to infants that is recognised by listeners from different cultural and musical backgrounds. In a further study, Trehub, Unyk and Trainor (1993b) low-pass filtered 30 pairs of lullabies matched with non-lullaby comparison songs with a similar tempo, culture and style. They found that listeners' rates of correctly selecting the lullabies was similar for the low-pass filtered and the original versions (about 67% correct selections). Songs with fewer changes in pitch direction (so simpler contours, and more descending pitch contours) were most likely to be selected as lullabies (Unyk, Trehub, Trainor & Schellenberg, 1992). These falling contours and restricted pitch range are also identified in soothing styles of IDS (Fernald & Simon, 1984).

Bergeson and Trehub (2002) recorded 19 mothers speaking and singing to their 4–7-month-old infants across two recording sessions, spaced at least one week apart. The mothers sang the same song (of their choice) each time. They were asked to speak naturally to their child at the first session and then prompted to repeat some of the same utterances at the second session by a list of their stereotyped phrases from the first. The findings indicate that mothers reproduce their song pitch and tempo highly accurately compared to their

speech pitch and tempo across the two recording sessions. In contrast to the variance in pitch and tempo, their speech rhythms were remarkably stable across the two datasets. The authors suggest that the consistent pitch and tempo in the song performances promote their soothing qualities, create social bonds and perhaps direct infants to attend to certain words. The differences in speech pitch and tempo may arise from the purpose of attracting infants' attention (Fernald, 1992), which could be facilitated with a more noticeable change in pitch, whereas the soothing purpose of singing could be facilitated by performing a familiar song exactly as an infant expects to avoid additional arousal (Bergeson & Trehub, 2002).

Research consistently shows that infants with functional hearing (Cooper & Aslin, 1990; Frank et al., 2020) and those with hearing aids (Wang, Bergeson & Houston, 2018; Robertson, von Hapsburg & Hay, 2013) prefer IDS over ADS, ID song over AD song (Unyk et al., 1992), and ID song over IDS (Trehub & Nakata, 2001; Tsang et al., 2017). The IDS over ADS preference is apparently present from birth, strengthening with age and language exposure (Cooper & Aslin, 1990; Byers-Heinlein et al., 2020), but it may diminish around 13 months (Outters, Schreiner, Behne & Mani, 2020). The magnitude of IDS preference varies across studies, with a meta-analysis of 34 studies (overall n = 840 infants aged 2–270 days, mean age = 138 days) finding a large Cohen's d of 0.67 (95% CI = 0.57, 0.76), z = 3.75, p = .0002 (Dunst, Gorman & Hamby, 2012). However, a comparative meta-analysis of 12 language development meta-analyses found that studies of IDS preference (and other aspects of language development) are routinely underpowered, with p-values used as a basis for interpreting non-significant findings as an 'absence of effect', and other questionable and limiting methodological practices (Bergmann et al., 2018).

Given a perceived lack of standardised testing and reporting procedures in this area, and thus to test the robustness of the IDS preference finding, a recent large-scale multinational study with 2,329 infants (aged from 92 to 456 days, mean age = 292 days) across 67 labs ran trials of infants' preferences for IDS or ADS using the three predominant

'looking-time' methods (Frank et al., 2020). The three methods used to test looking-time (as a proxy measure for infant preference in either IDS or ADS) were central-fixation on a visual stimulus (20 studies), head-turn preference procedures (HPP; 21 studies), and eye-tracking (30 studies). The authors conducted a meta-analysis and report a much smaller overall effect size for infants' preference for IDS over ADS than previously reported, with Cohen's $d = 0.35$ [95% CI = 0.29, 0.42], $z = 10.67$, $p < .001$. When comparing the findings by method, HPP studies returned the largest effect size, then central fixation and finally eye-tracking. However, the authors caution that these data do not indicate that HPP is better suited for testing all research questions since methods were not randomly assigned to labs, and labs with more experience in language acquisition studies may have systematically chosen HPP, thus introducing bias. The data also indicate an age effect, with older infants preferring IDS to ADS, which the authors interpret as potential methodological artefacts of older infants showing more measurable responses, or the stimuli being better suited for older infants, and also that preference for IDS might be modulated by maturation and experience. Overall, despite the overall effect size being much smaller than reported in Dunst et al. (2012) or prior related reviews (Bergmann et al., 2018), there is still a fairly robust effect of IDS over ADS on infant preference.

Having established that infants demonstrate a preference for IDS, the next consideration is of the potential functions of IDS and ID song. A meta-analysis of the relations between the prosodic aspects of IDS and infant outcomes (Spinelli, Fasolo & Mesman, 2017) found that improved attentional, pre-linguistic and linguistic outcomes were associated with prosodic values typical of IDS. The prevailing assumption is that IDS facilitates language acquisition (Fernald, 1992) because it evokes and maintains infant attention on the caregiver (Fernald et al., 1989; Garnica, 1977), on their face and – increasingly towards age one – on their mouth (Alviar, Sahoo, Edwards, Jones, Klin & Lense, 2023; Lewkowicz & Hansen-Tif, 2012; Tenenbaum et al., 2015), and on the object of

shared attention (a key factor in Tomasello's 2003 usage-based account of L1 acquisition; Dominey & Dodane, 2004). IDS also elicits more neural activity compared to ADS (Zangl & Mills, 2007), which leads to the interesting question of how the speech signal is encoded in the brain, and what prosody's role in encoding is that would lead it to be prioritised in the way adults often address children.

In one possible solution to this encoding question, Kreiner and Eviatar (2014) suggest that prosody is the link between the abstract nature of language and the physical functions of the brain. They posit that brain activity resonates to the rhythm, stress, pitch and intonation (i.e., the prosodic) features of the acoustic stream. The brain maps prosodic features of language with corresponding patterns of neural activation, representing syntax in the brain: the 'embodiment' of syntax. Introducing models of 'structure grounding', Kreiner and Eviatar (2014) posit that language is grounded through reactivating brain states that are associated and encoded with motor, perceptual and internal experience at the time of first experiencing an input. They give the concrete example of eating an apple where brain states that are associated with the senses, eating movements and internal changes (e.g., blood sugar rising) are captured in a multimodal representation of the experience. Just thinking about eating an apple can then reactivate or simulate that earlier experience in a symbolic abstraction of the experience that was embodied through the perceptual, motor and introspective states. This is why our mouth waters when we think about eating something we enjoy.

Returning to prosody, then, according to this idea of embodiment, linguistic experience is encoded in a series of physical events associated with the brain's perception of rhythm, stress, pitch, and so on, in the acoustic signal. Abstract mental representation using language in place of physical objects or context allows humans to use language symbolically to talk about objects or events that are not currently present: communication is dissociated from the physical context, and this symbolic use of language is a uniquely human capacity

(see Tomasello, 2003, for discussion). It could be that ID speech and song are vehicles for a multimodal embodiment of language where positive affect is mutually elicited between the caregiver and infant (Sharman et al., 2023), leading to a relaxed state (Bainbridge et al., 2021) in which the infant is more perceptive of speech sounds from the acoustic signal (François et al., 2018) and encodes these prosodic cues in neural networks (Kreiner & Eviatar, 2014), leading to language development over the first two years through frequent repetition (Franco et al., 2021).

Lastly, in response to the 'logical problem' of language acquisition (see section 2.3.1.3.3), MacWhinney (2004) posits a multi-process solution that asserts input's crucial role in children's language acquisition because input provides positive data of how language is structured and used, which is essential for learning. Notably, MacWhinney (2004: 911) includes parental recasts as positive data (which are part of how children learn to avoid overgeneralisations, one of the nativist concerns about the input being the sole source of linguistic information), as well as listing "elicited repetition, choral recitation of stories, interaction with siblings, or games." Although MacWhinney does not mention them specifically, positive data could also be presented through songs.

In summary, songs, and in particular ID songs, may form part of the richly informative ID input whose prosody plays a crucial role in infants' early linguistic development, which is part of their holistic socio-emotional and neural development. Parents' performances of ID songs, infants' preferred source of input compared to AD songs or even IDS, are a stable form of linguistic input that varies very little across time. Such input, far from being 'degenerative' (Chomsky, 1965), provides an intuitively tailored and multimodal source of cues that potentially lead to superior encoding of the acoustic linguistic signal in infants' developing neural networks. It remains to be seen whether similar prosodic input cues are exploited at later stages of childhood in FL instructed contexts and to what extent these early L1 acquisition mechanisms remain online in input-limited FL education.

2.4 Summary and conclusion of literature review

As we have seen, there is a wide body of research that includes singing songs as a positive source of linguistic input to young children. Songs are written into FL teaching curricula in the UK and assumed to have positive effects on the linguistic development of YLLs in instructed contexts in government curricula (DfE 2013a) and programmes of work for primary FL (DfES, 2005a, 2005b). There is also evidence dating back to the Middle Ages of educators' stable belief that singing and oral language development in classrooms go hand in hand. One could, based on this review of the literature so far, begin developing a primary school intervention study to investigate the linguistic effects of singing songs with YLLs to contribute to the ongoing discussion of their merits.

Furthermore, although teachers appear to 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, it appears that 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.

However, the review of the literature presented in my thesis thus far is 'narrative' (Gough, Oliver & Thomas, 2012) rather than systematic, in that I have developed a story about how songs are used in teaching FL to YLLs, and hand-selected evidence to support the

thread of my narrative rather than specifying by which methods I would include evidence. This approach to reviewing the literature is not exhaustive, and it is likely that despite my best efforts, I have failed to uncover all the relevant evidence from FL classrooms that would inform the next stages of this study. It would be wasteful (Chalmers et al., 2014; Isaacs & Chalmers, 2023) to embark on a primary research study without first ascertaining that the answer has not already been provided by existing research. For this purpose, Phase 1 of this doctoral research presents a systematic review of the literature (Petticrew & Roberts, 2006). The review, reported in the next chapter, contributes useful substantive evidence of what is already known about songs' effectiveness as pedagogical tools for teaching YLLs and serves as a solid foundation for primary research to build on.

Chapter 3

Systematic review

3.1 Introduction

This chapter presents my systematic review of intervention research that investigates the effects of songs and singing on young learners' linguistic outcomes in foreign language lessons in school. A systematic review is a primary research study that is structured, transparent and replicable and reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement (Moher et al., 2009). The chapter will follow the structure therein, presenting the objectives of the review, the methods used to gather the included studies, the results, discussion of the findings and their implications. The results form a narrative synthesis of the included studies, rather than a meta-analysis, since there was such heterogeneity of participants, methodologies and reporting standards that it would be unreliable to synthesise the collected results or calculate effect sizes beyond the individual studies. The discussion looks at implications for the research field and for teaching practice. I then discuss how the review contributed to the design of Phase 2 of this doctoral research – the intervention study.

3.1.1 Why do a systematic review?

I had never heard of a systematic review before I began my doctoral study. Yet it seems to me now that not conducting a *systematic* review, as opposed to a *narrative* review (Gough et al., 2012) guided by my own tastes and biases for where the literature took me, would have been akin to sticking a pin in the map before going travelling and then just seeing what looked like a good place to stay on arrival in the dark, or basing this decision on recommendations from friends, the travel guide I skim read at the airport, or something I saw in a movie. A narrative review can lead one astray: whilst we may believe a well-rounded

and diligently researched presentation of the 'state of the art' digest of the topic has been achieved, we are nonetheless prey to our own biases, forgetfulness, distractions, readily available resources, and networks. Potentially this means a narrative review has, either consciously or unconsciously, selectively included studies that help us tell a story we set out to narrate, which can distort reality by leaving out elements that do not 'fit' our narrative.

On the other hand, systematic reviews set out to be trustworthy, replicable and systematically methodical in assembling and evaluating the collected literature (Gough et al., 2012; Petticrew & Roberts, 2006). The researcher is accountable for the processes by which they carried out the literature searches, and transparently reports the process to the extent that future researchers can replicate their methods exactly to update the search and add any new research findings to our collective understanding of the substantive topic. Searches may include only peer-reviewed papers, but a more exhaustive search will seek out theses and other 'grey' literature (Schöpfel, 2010) such as conference proceedings, or by writing to researchers in the field to ask if they know of any unpublished work that meets the review's eligibility criteria.

By aiming to present a complete review of the relevant work to the review's research questions, the researcher can proceed with their next steps with more certainty that they are not wasting time and resources going over old ground, nor missing a key opportunity to break new ground and, importantly, driving the field forward with more nuanced research questions. For practitioners and policy makers, a systematic review provides more trustworthy and complete information on which to base their real-world decisions, many of which will have real risk and real-life consequences attached. Reducing bias in the way we search for, include, evaluate and synthesise the literature is therefore of paramount importance to real people, whether that is in the case of providing a choice of medical treatments or the perhaps less life-and-death but no less important choice of pedagogical approaches in foreign language learning contexts.

3.1.2 The focus of this systematic review

Before embarking on an intervention to investigate the effect of using songs on children's linguistic outcomes, I conducted this systematic review of existing intervention studies to gain as much insight as possible into the existing empirical evidence base, with a view to informing the design of my intervention. A narrative review of the literature, as presented in Chapter 2, identified key theoretical and empirical avenues of exploration that suggest singing songs with young language learners is a popular approach with many languages teachers, and potentially beneficial for their linguistic outcomes. However, there did not appear to be a univocal message emerging about what songs' effects on linguistic outcomes are, exactly which linguistic outcomes are likely to benefit from a teaching approach that uses songs, or (regarding pedagogical recommendations) how best to present songs to achieve particular language learning goals. Therefore, by taking a systematic approach to reviewing the intervention literature, the extent of our collective knowledge about this topic can be established. The process of locating, appraising and synthesising research that compares singing to other teaching approaches and measures the effects on linguistic outcomes highlights gaps that could then usefully inform Phase 2's intervention study.

3.2 Structured summary

3.2.1 Background

Songs are popular resources with teachers of YLLs. In addition to important socioemotional and developmental outcomes, a common assumption is that songs will help support learning the target language. Whilst there is clear anecdotal support from practitioners for using songs to achieve linguistic (amongst other) FL outcomes, there appears to be 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).

3.2.2 Objectives

This review had two aims. Firstly, to ascertain 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. Secondly, to assess what can be reliably concluded from included intervention research about the effects of using songs with young language learners on substantive foreign language learning outcomes.

3.2.3 Methods

Studies considered for inclusion: intervention studies published in any language on any date conducted in preschool, primary or secondary schools with children aged 2–18 that assessed the effects of using songs, chanting or nursery rhymes on linguistic outcomes. 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).

Information sources: The following English databases were searched: ProQuest Education Collection (including ERIC), British Education Index EBSCO; Education Abstracts (H.W. Wilson), ProQuest Linguistics Collection (including LLBA); MLA International Bibliography, PsychInfo, Web of Science, Scopus, ProQuest Dissertations & Theses Global; OpenGREY; EthOS. Additionally, the following German, French and Spanish databases were searched: Fachportal Pädagogik; PsynDEX; Humboldt University Berlin; Center for Research Libraries Global Resources Network (CRL); Cairn.info; SUDOC; Pascal-Francis; CRL; theses.fr; TESEO education.gob.es.

Risk of bias: The Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018; Pluye, Gagnon, Griffiths & Johnson-Lafleur, 2009) was used to assess the Risk of Bias of included studies.

3.2.4 Results

After screening, 60 intervention studies from 23 countries published between 1978 and 2021 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. There are 43 peer-reviewed articles and 17 theses *(n* = 3 master's, *n* = 14 doctoral). 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. Study duration ranged from one hour to two years. 83% (*n* = 50) studies are published in English, followed by Korean (5), Spanish (4), and French (1). The number of participants ranged from 5 to 573, with a median of 56 participants. A narrative synthesis was conducted since statistical synthesis of the findings was not feasible given the heterogeneity of methods, participants and outcome measures of included works. Most studies report songs' positive effects on their measures, but the cumulative weight of evidence is limited. Based on the MMAT quality appraisal, of the 60 included studies, three received 'strong', 14 'moderate', and 43 'limited' global weight of evidence ratings. No overall conclusions about songs' effects on linguistic outcomes could therefore be drawn.

3.2.5 Discussion

While most of the assembled literature made positive causal claims about the relationship between singing songs and linguistic outcomes, a majority were not appropriately designed to support these claims. The 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. Teachers contemplating using songs as a tool for teaching FL cannot therefore draw upon reliable evidence to inform their pedagogical choices. 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.

3.3 Aims of the systematic review

My aim in conducting this systematic review was to describe 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 hoped that by assembling, describing, and evaluating intervention research, this review will provide important information about the use of songs in FL so that teachers can make informed decisions about their practice and its likely effects on their students' linguistic outcomes. The review findings were used to inform the intervention design of Phase 2 of this doctoral research.

3.3.1 Review questions

This review addresses the following research questions:

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?

3.4 Methodology

3.4.1 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, Albert & Piniel, 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).

3.4.2 Eligibility criteria

Table 3.1 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.

Table 3.1 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.	This review focuses on the linguistic outcomes of using songs as pedagogical tools, thus the

	Exclude 5: Studies where musical instruments are the intervention focus, not singing, chanting or nursery rhymes.	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.

3.4.3 Information sources

Table 3.2 shows the consulted databases in education, linguistics, psychology, and multidisciplinary research. I 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 3.2 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

3.4.4 Search strategy

The main search strategy for ProQuest (see Table 3.3) 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. Searches were limited to the musical nature of the intervention, not type of study design, and instead design was applied during the selection process.

Table 3.3 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*)

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 I captured maximum relevant results per language (see examples in Appendix A).

3.4.4.1 Citation chaining

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

3.4.5 Selection Process

Following deduplication, I screened titles and abstracts using Rayyan software (Ouzzani, et al., 2016). Records clearly violating one or more inclusion criteria were excluded. A fellow doctoral researcher, blind to my decisions, independently screened a randomly selected 10% sample (n = 184) of titles and abstracts. We compared decisions and discussed discrepancies

(n = 5, κ = 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, I sought them through interlibrary loan or emailing authors. I screened 89 full texts, excluding any that violated inclusion criteria. A Korean applied linguistics colleague screened the Korean papers. Ambiguous inclusion decisions not included in the prior collaborative 10% screening were discussed, and agreement reached.

3.4.6 Data collection process

Before completing final searches, I created a data extraction form (see Appendix B), adapted for relevance to the focus of the review from the principles laid out in the Cochrane Good Practice Guide (Cochrane Effective Practice and Organisation of Care, 2017) and Boland, Cherry and Dickson (2017). I piloted the form on two included studies (Davis & Fan, 2016; Chou, 2014), ensuring it 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 I completed data extraction for 60 papers, a doctoral colleague 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.

3.4.7 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 et al., (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.

3.4.8 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, 2024). 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. Many protocols for systematic reviews in language education registered on IDESR have adopted the tool, and it is 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 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, Altman & Gøtzsche, 2011) and Cochrane ROBINS-I (Sterne, Hernán & Reeves, 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 Chapter 2), an objective and rigorous tool such as the MMAT helps to ensure review conclusions reflect the trustworthiness of included evidence.

My doctoral colleague independently appraised six included studies (two randomly chosen theses, four full reports) after I 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. We discussed the interpretation of those items and resolved disagreements through discussion.

3.4.9 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), a narrative synthesis was conducted 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 3.5.4 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

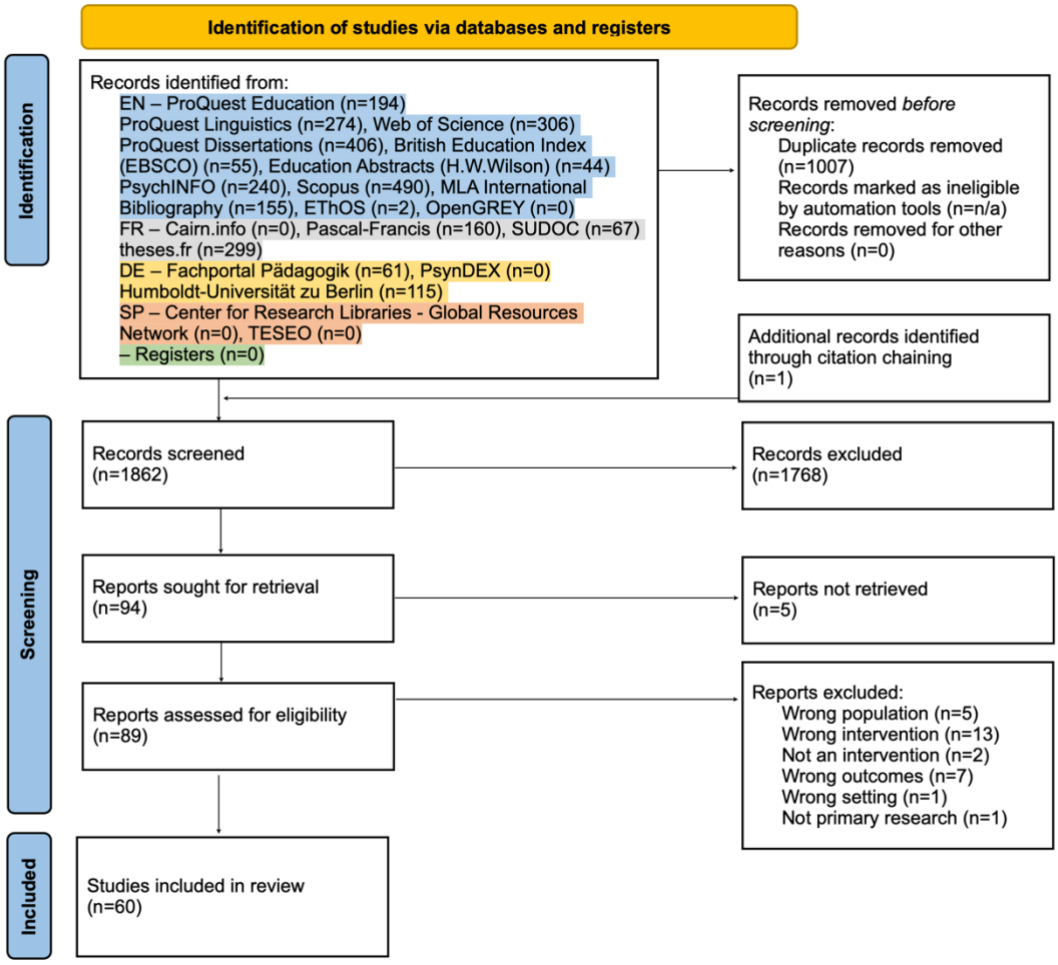
mixed findings are coloured yellow; significant differences in favour of the control group are coloured pink. Combined with the MMAT colour-coding (green = strong, yellow = moderate, pink = limited trustworthiness ratings), it is possible to visualise positive or negative findings and the trustworthiness of studies within each category.

3.5 Results

3.5.1 Study selection

Figure 3.1 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 (Joyce, 2011; Lehman, 2019; Crosswhite, 1996).

Figure 3.1 PRISMA 2020 flow diagram of study selection process



3.5.2 Study characteristics

Table 3.4 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.

Table 3.4 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/posttest	Romania	34	Secondary	15	Grammar; attitudes	Grammatical knowledge test
3. Allen-Tamai (2000)	PhD	Non-equivalent groups pre/posttest	Japan	62	Preschool	11	Phonological awareness	Rhyme awareness
4. Alley (1988)	PhD	Non-equivalent groups pre/posttest	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/posttest	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/posttest	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/posttest	Malaysia	573	Primary	2 school years	Four skills	Speaking, listening, reading, dictation
12. Busse, Hennies, Kreutz & Roden (2021)	JA	Non-equivalent groups pre/posttest	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/posttest	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/posttest	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/posttest	France	20	Primary	7	Speaking	Pronunciation (L2 vowels)
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/posttest	Taiwan	72	Primary	5x 100-minute 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/posttest	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, posttest 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/posttest	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; reading; attitudes	Listening; vocabulary/ grammar; reading (not defined further)
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/posttest	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/posttest	Taiwan	47	Preschool	6–8	Vocabulary; speaking	Pronunciation and oral spelling of colours
32. Jarvis (2013)	JA	Non-equivalent groups pre/posttest	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/posttest	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/posttest	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 hour	Vocabulary	Translate English phrases containing passé composé / imperfect verbs into French equivalent; translate weekdays
39. Leśniewska & Pichette (2016)	JA	Single group posttest only	Canada	24	Preschool	4	Vocabulary	PPVT
40. Lowe (1995)	PhD	Non-equivalent groups pre/posttest	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 groups crossover	UK	59	Secondary	4	Vocabulary; grammar; attitudes	Cloze test of song lyrics; translation French > English

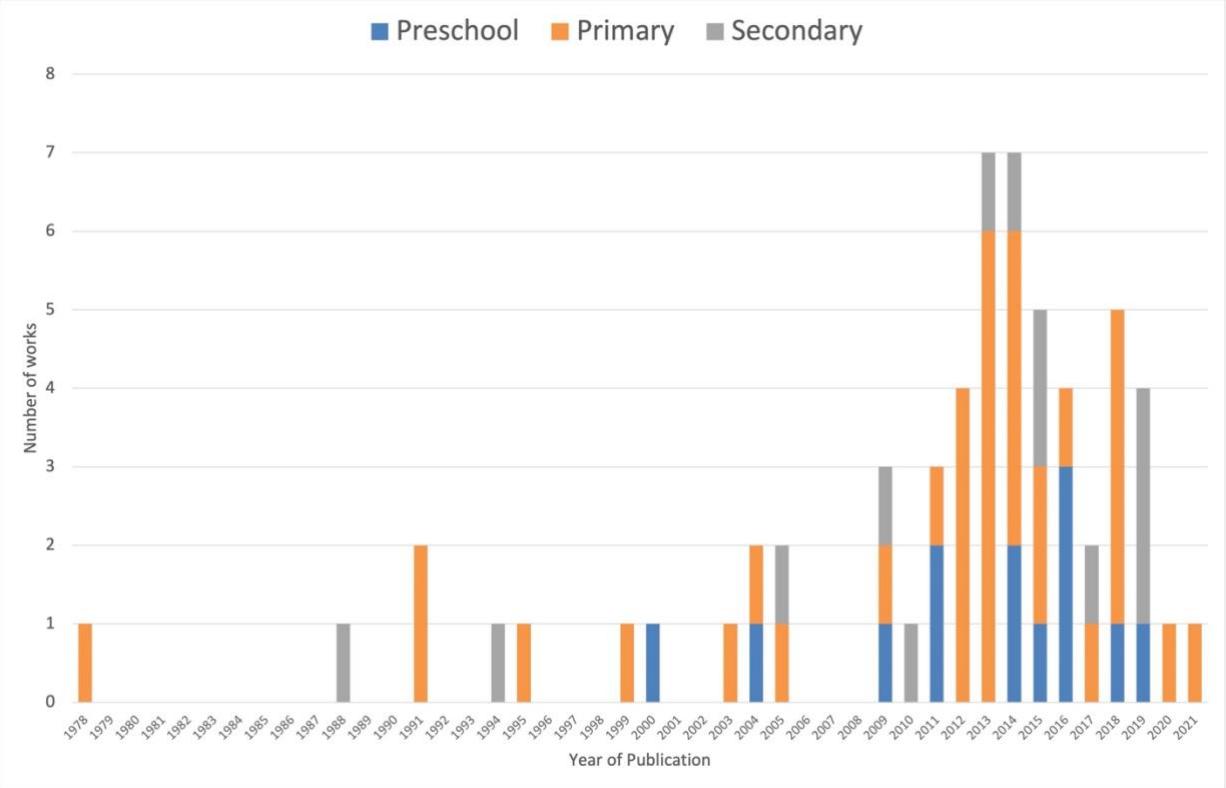
42. Luo (2019)	JA	Single group pre/post	China	50	Secondary	3	Vocabulary; attitudes	Use target words in a sentence; Chinese > English word translation
43. Ma (2004)	JA	Single group pre/post	Korea	48	Preschool	4	Vocabulary; story recall	Picture vocabulary test: point (receptive) and label (productive); child prompted to complete sentences by reading/singing along with story
44. Madani & Nasrabadi (2016)	JA	Non-equivalent groups pre/posttest	Iran	112	Preschool	1 month	Vocabulary	Vocabulary learning/retention
45. Mamdouh (2017)	JA	Non-equivalent groups pre/posttest	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/posttest	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/posttest	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/posttest	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 minute) silent reading fluency & word identification/segmentation
58. Wang (2005)	MSc	Non-equivalent groups pre/posttest	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
60. Zhaku-Kondri (2014)	JA	Cluster RCT	Macedonia	57	Primary	8	Vocabulary; grammar; attitudes	Grammar (verb tenses) in pre/posttests; vocabulary in posttest

3.5.2.1 Publication details

Figure 3.2 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 1979–1987; since 2009, a further 47 eligible studies were published, 23 of these from 2013–2016. There are 43 peer-reviewed articles and 17 theses (n = 3 master's, n = 14 doctoral).

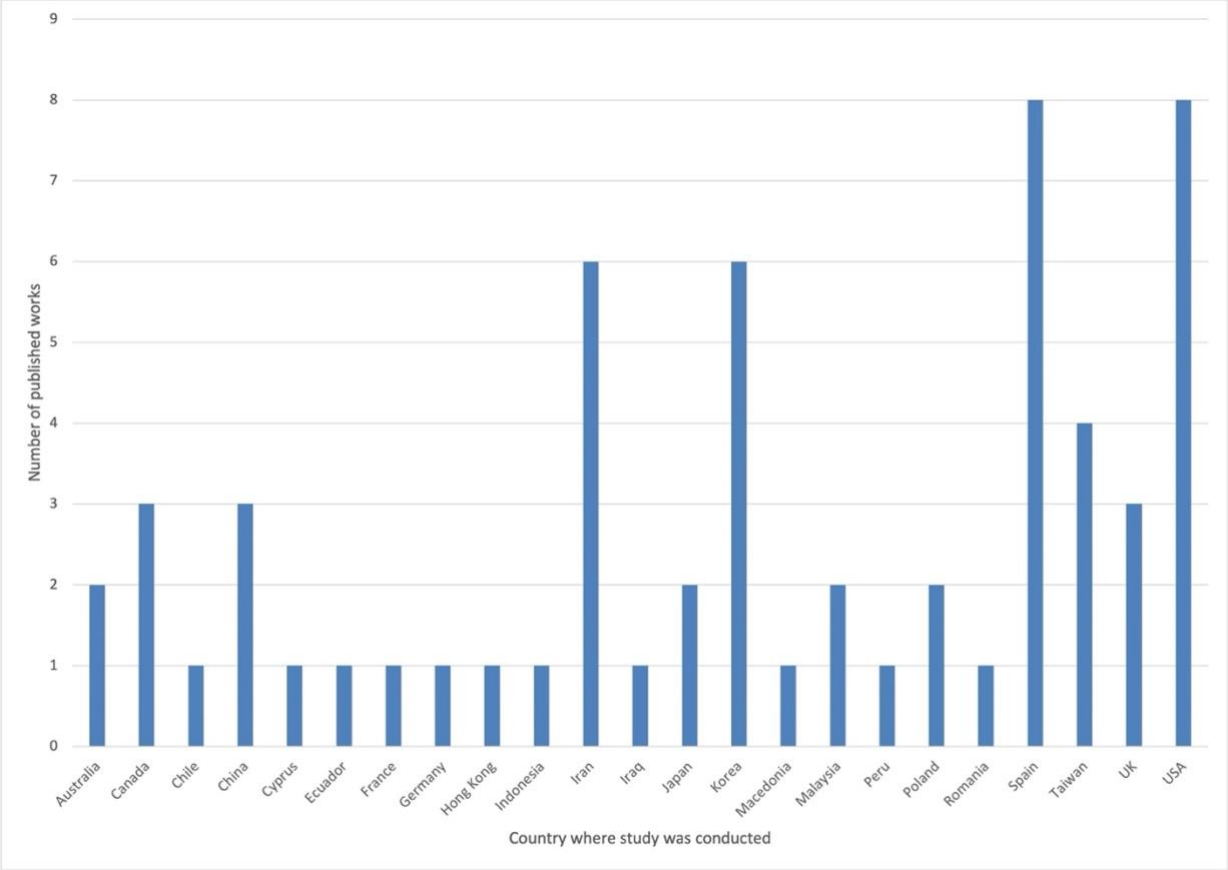
Figure 3.2 Number of included studies by publication year and educational context



3.5.2.2 Geographic context

Studies included in this review were conducted in 23 countries (Figure 3.3), from all continents except Africa. 83% (n = 50) studies are published in English, followed by Korean (5), Spanish (4), and French (1).

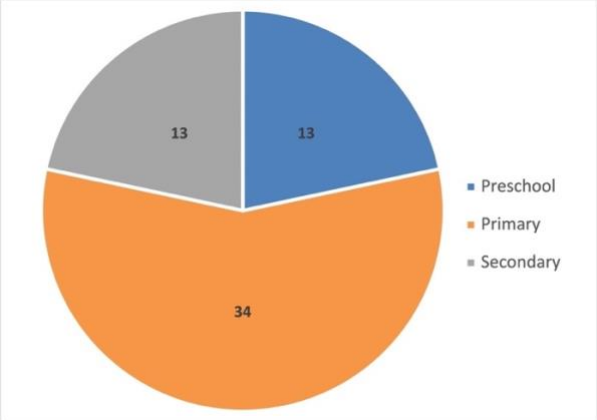
Figure 3.3 Number of published works by geographic context



3.5.2.3 Instructional context

Figure 3.4 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 2012–2014 (see Figure 3.2).

Figure 3.4 Instructional context of studies



3.5.2.4 Study design

Table 3.5 summarises included studies' designs.

Table 3.5 Summary of study designs

Study design	**No. of Works**	**Study ID**[13]
Non-equivalent groups pre/posttest	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/posttest	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, posttest only	1	24
Single group, posttest only	1	39

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

3.5.2.5 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].

3.5.2.6 Allocation strategy

Table 3.6 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 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/posttest design.

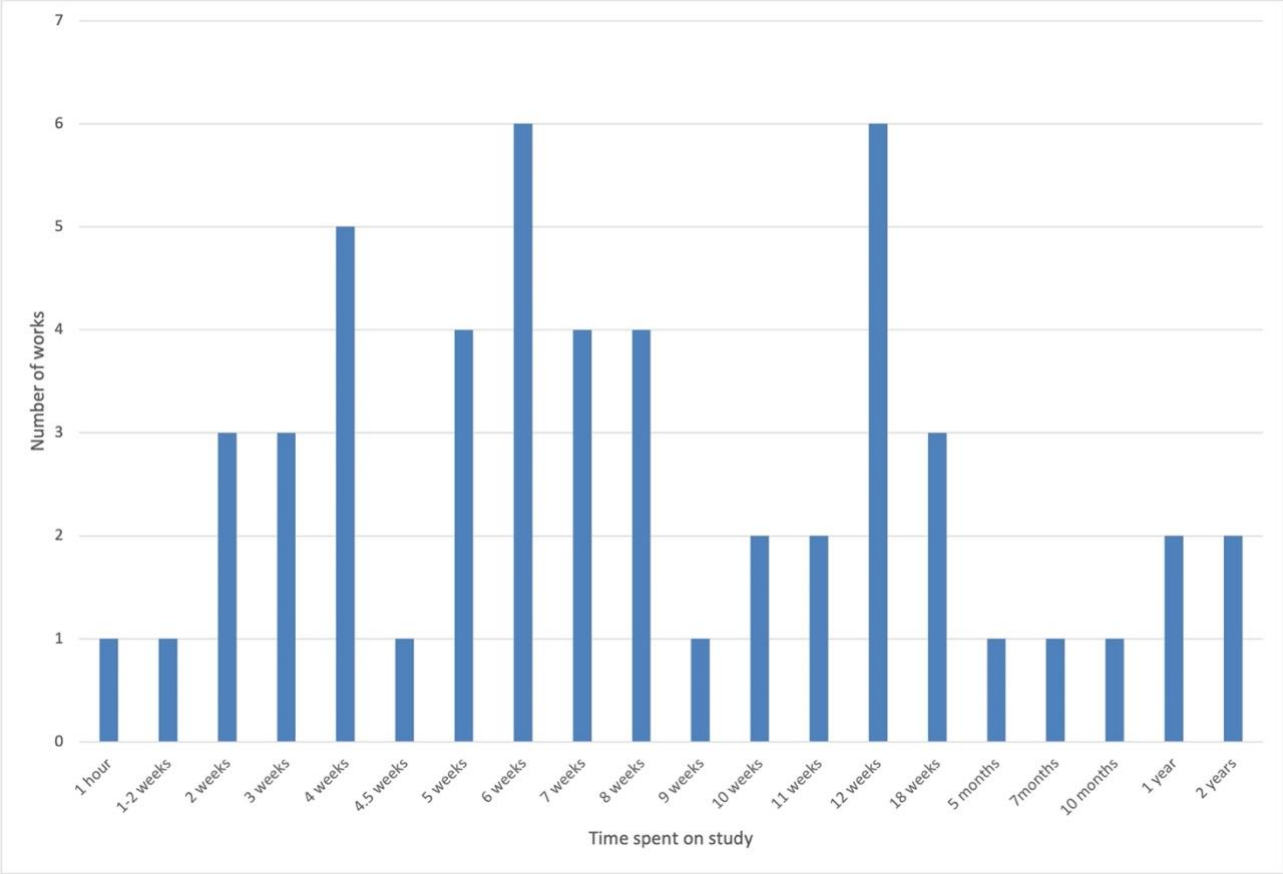
Table 3.6 Allocation strategy

	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
CLUSTER	Intact classes (not randomly assigned)	4	2, 4, 41, 43
CLUSTER	Random allocation of intact classes by drawing class names from a hat (first to be drawn assigned to music condition)	1	37
CLUSTER	Random allocation of intact classes (no strategy reported)	6	8, 17, 18, 26, 53, 60
INDIVIDUAL	Random allocation at individual level (strategy not reported)	7	5, 6, 14, 27, 28, 38, 59
INDIVIDUAL	Individuals matched by pretest scores and randomly assigned to four groups, then groups assigned to conditions by shuffling papers with names of the groups on	1	48
INDIVIDUAL	Matched by pretest scores and randomly assigned to conditions by flipping a coin	1	36
INDIVIDUAL	Matched by grade level, school and gender and assigned to conditions (allocation strategy not reported)	1	54
INDIVIDUAL	Children's names alphabetised within their groups, assigned numbers, then odd numbers assigned to comparison and even to treatments (i.e., alternation)	1	24
INDIVIDUAL	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

3.5.2.7 Study duration

Study duration ranged from one hour to two years. Three papers[32,50,53] fail to report duration. Figure 3.5 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.

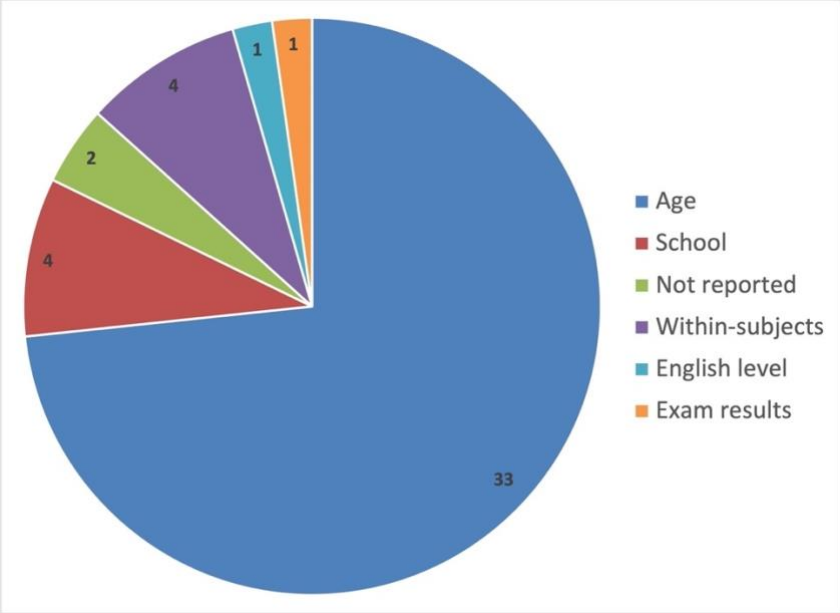
Figure 3.5 Study duration



3.5.2.8 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 Figure 3.6 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].

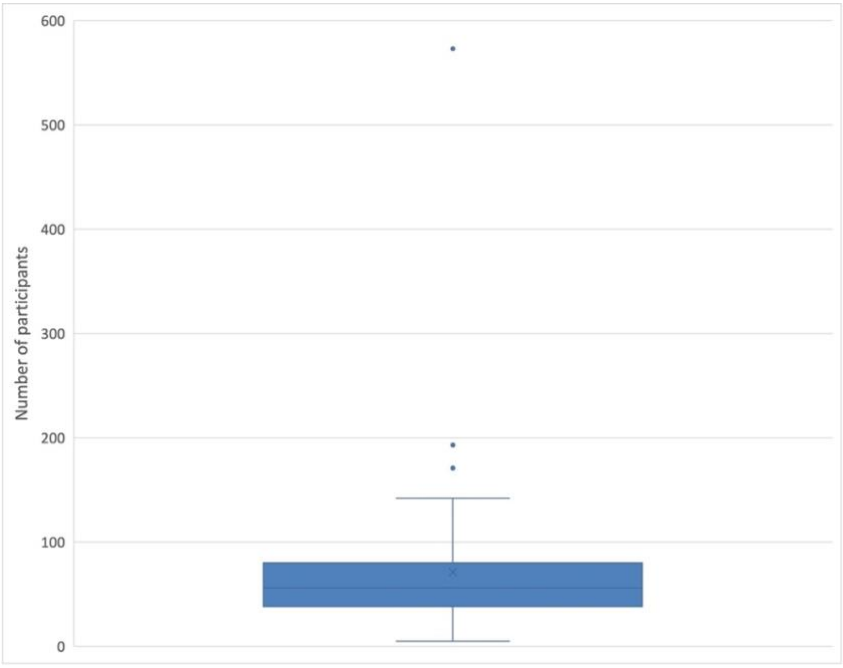
Figure 3.6 How comparison groups are matched



3.5.2.9 Sample size

Figure 3.7 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 Figure 7 presents.

Figure 3.7 Sample size



3.5.3 General reported outcomes

Figure 3.8 illustrates the outcomes reported by included studies: vocabulary, grammar, four skills (listening, reading, writing, speaking), attitudes, or other. Figure 3.9 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 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.

Figure 3.8 Outcome type by study

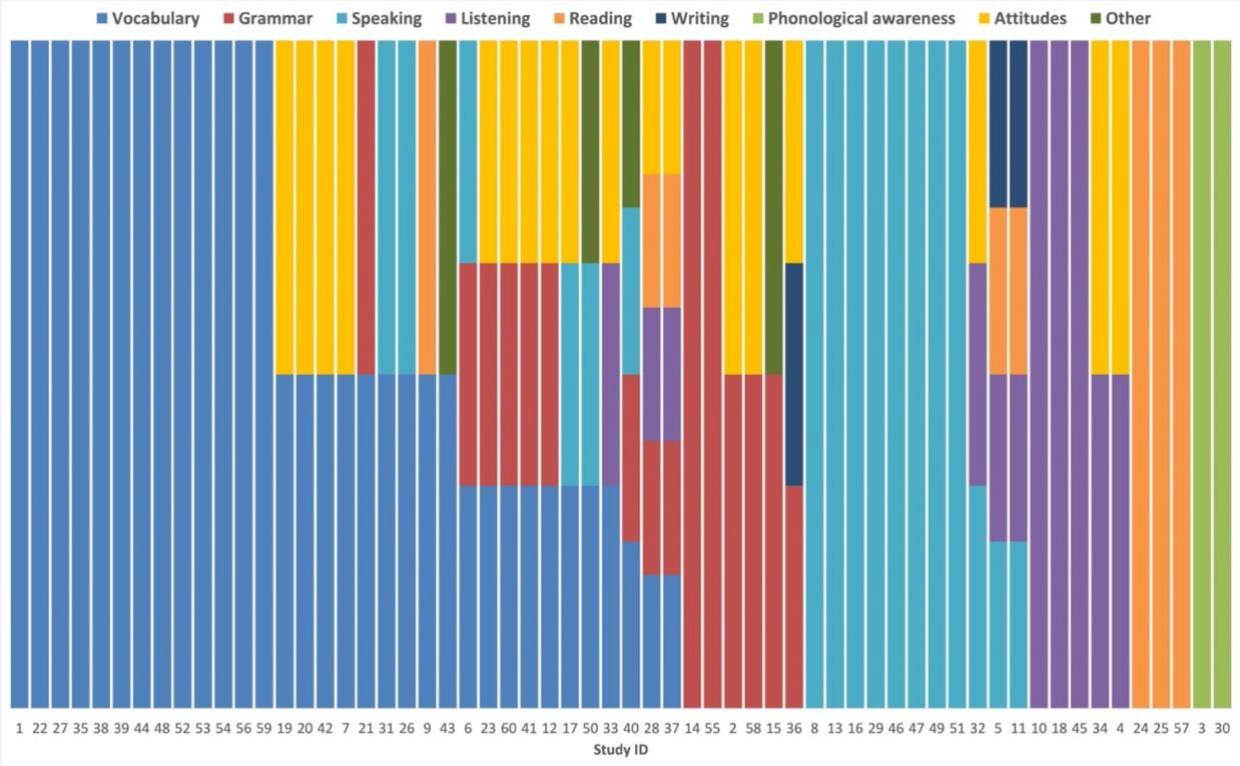
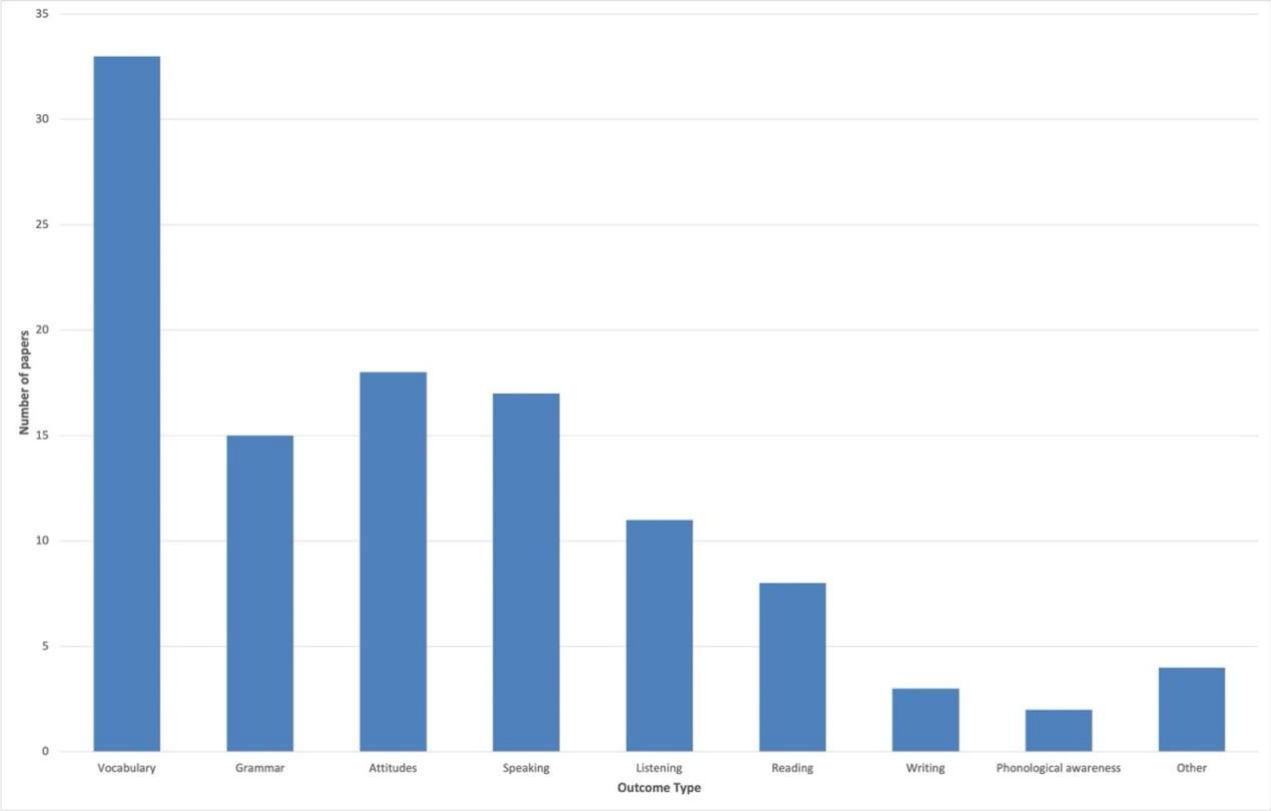


Figure 3.9 Outcome frequency



3.5.4 Specific reported measures

3.5.4.1 Vocabulary measures

The largest group of studies (n = 33) report vocabulary outcomes. 13 studies measure receptive vocabulary (summarised in Table 3.7) and 13 measure productive vocabulary (Table 3.8). 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 posttest. 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 3.8 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 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 3.7 Studies reporting receptive vocabulary measures **unclear whether standardised PPVT was used*

	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 & Gómez 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 (2016)	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 posttest 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

Table 3.8 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 & Gómez 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 128erano128/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

3.5.4.2 Grammar measures

Table 3.9 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 intact classes where the sample is $n = 2$.

Table 3.9 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 & Sobouti (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 posttest, 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 posttest; 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 posttest (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

3.5.4.3 Speaking measures

17 studies measure L2 speaking skills, five[5,11,13,32,50] not clearly reporting how outcomes were measured. Table 3.10 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 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.

Table 3.10 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 & Sobouti (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 posttest 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 & Shahrokhi (2014)	Posttest of pronunciation, intonation, stress recognition (each marked out of 10). Recordings of posttest 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

3.5.4.4 Listening measures

11 included studies investigate the effect of singing interventions on L2 listening skills but nine[5,10,11,18,28,32,33,34,45] fail to report their measures in enough detail to synthesise. Table 3.11 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 pretests, 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.

Table 3.11 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 posttest. Treatment groups scored significantly higher than no treatment groups in posttest. 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 posttest differed (text presented as song in units and not in posttest, 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.

3.5.4.5 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 3.12 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 children from Sinti and Roma backgrounds, 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-grapheme correspondences and decoding skills during timed silent reading tests. It reported positive findings but was a single-group pre/posttest 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.

Table 3.12 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 posttest, but it does not report if the treatment groups' mean scores were significantly different from each other.	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.

			Naming task: treatment groups outperformed controls. Tamazight (L2 Spanish) learners in the music group significantly outperformed Tamazight learners in the control group.	
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 children from Sinti and Roma backgrounds 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 posttest, 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

3.5.4.6 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 3.13 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.

Table 3.13 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.

3.5.5 Risk of bias

Risk of bias (RoB) assessment results for each included study are summarised in Table 3.14. 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 3.5.4 above.

3.5.5.1 Cumulative confidence across studies

Of the 60 included studies, three received 'strong', 14 'moderate', and 43 'limited' global weight of evidence ratings. As Figure 3.10 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 3.6).

Most studies report songs' positive effects on their measures, but the cumulative weight of evidence is limited, as Figure 3.11 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.

Figure 3.10 Global weight of evidence ratings (MMAT)

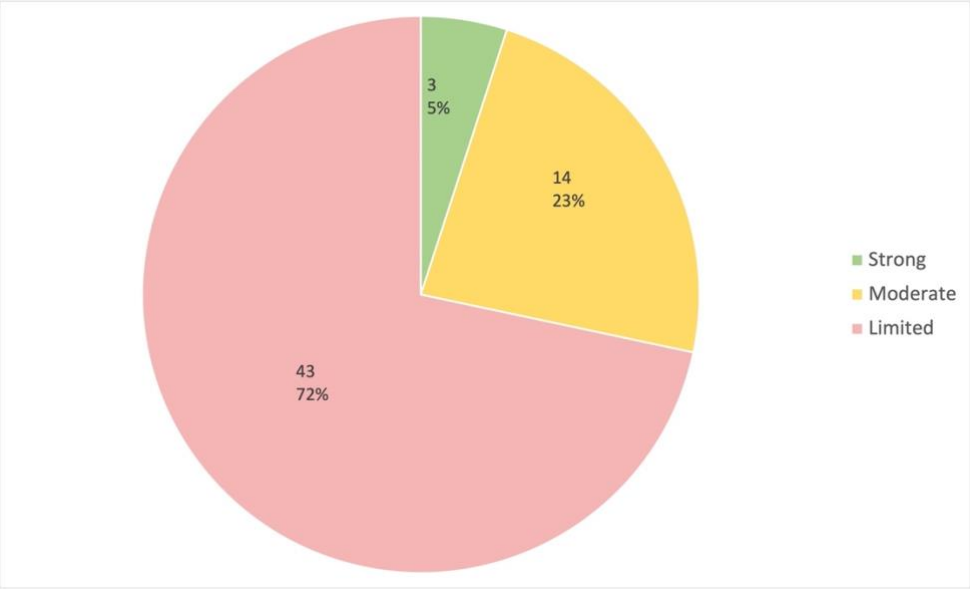


Table 3.14 Risk of bias of individual studies

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

Figure 3.11 Reported effect of singing and weight of evidence

3.6 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 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 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). 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 & 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)[20], 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).

3.6.1 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, my 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.

3.7 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 this review includes any intervention research design where linguistic outcomes were measured, due consideration of the limitations of these designs is needed. I 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, I 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 review 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.

3.8 Informing Phase 2 of this doctoral project

This systematic review's aim was to ascertain the extent and nature of intervention research investigating the use of songs with YLLs, and thereby identify where Phase 2's intervention study would helpfully fill a 'gap' in our evidence base about songs' effects on linguistic outcomes. Despite the confident claims made many teacher-facing resources about using songs in various guises to achieve linguistic outcomes in school settings (Davanellos, 1999; Paquette & Rieg, 2016; Thain, 2010), it appears that very little robust empirical research has been conducted to compare the relative effects of song-based and alternative teaching approaches on well-defined measures of children's L2 progress or development. There have been very few randomised controlled trials which could reliably detect a causal effect, were there to be one present in an experiment. In terms of informing Phase 2 of this doctoral research, based on the findings from the review it would be justifiable to choose any linguistic outcome, and any experimental design involving songs, since there is not one 'gap' but rather a host of unanswered questions. We do not know with any certainty whether using songs as a

teaching approach has a detectable effect on children's linguistic outcomes or not. Furthermore, the theoretical foundations of much existing research also lack empirical evidence to support assumptions that using songs would have a differential effect on children's L2 development relative to other methods of presenting FL in classrooms. What this review points towards is threefold. Firstly, theoretical motivation needs to be better grounded in empirical evidence and theories should be testable, so that research can contribute to our theoretical understanding of the topic. Secondly, research questions need to build upon existing work rather than starting afresh with each study. Finally, experimental designs need to be more methodologically rigorous and transparently reported to permit quality appraisal, replication, and contribute effectively to the development of knowledge in this field. These three indicators are therefore what I chose to focus on in deciding the focus of my intervention study.

3.8.1 Deciding the focus of the intervention study

As uncovered during my systematic review of the literature, there are very few empirical studies that take a reliable experimental approach to exploring the topic of this thesis: *Investigating the effects of whole-class singing activities on linguistic outcomes of young language learners in English primary schools*. Only 34 of 60 included studies took place in primary schools, including all three studies found to be at low risk of bias. Few studies build upon the work of previous authors to develop increasingly nuanced research questions and situate their investigations within an empirically tested theoretical framework. The first aim of this thesis, then, was to build upon prior work and extend our knowledge about the topic of whether singing in primary school foreign language lessons has (relative to other methods of teaching languages) a measurable effect on children's substantive linguistic outcomes. Secondarily, the aim of this thesis is to contribute to our knowledge of how the prosodic bootstrapping hypothesis, an empirically tested theory of L1 acquisition (see discussion, Chapter 2), may apply in the second or foreign language classroom with young L2 learners.

In both of these aims, this thesis builds on the doctoral research of Campfield (2010), and ensuing journal publication by Campfield and Murphy (2014), which investigated the role of linguistic rhythm in L2 production skills with young learners of English in Polish primary schools, positioning that work within a prosodic bootstrapping theoretical framework. Campfield and Murphy (2014) presents a study where 80 Polish learners of English as a foreign language (average age 8;4 years) were randomly allocated to three groups. The experimental group recited English nursery rhymes (rhythmically salient input); the comparison group recited English stories (less rhythmically salient prose input that matched the experimental group for grammar and vocabulary items); and the control group continued with their regular English lessons, receiving no additional input. They found that the experimental group produced significantly more 'exact' imitations and was able to more accurately imitate longer sentences than the comparison and control groups on an elicited imitation task (EIT) measure of L2 spoken production. They argue that the reconstructive EIT taps into the structural complexity of participants' L2 through assessing their spoken L2 production. There are a number of reasons to build on this study and reconstruct it the UK educational context, where exposure to target foreign languages is limited. Firstly, it is one of the few studies in the field where participants are randomly allocated to conditions, and as such it provides a more reliable contribution to our knowledge about the effect of linguistic rhythm on young L2 learners' spoken production. The study included nursery rhymes and stories, but not songs, as the experiment and comparison conditions, and it would be straightforward to add a song condition to the existing experimental design, thereby testing whether adding melody as well as prosody to the input would produce replicable effects.

Also, Campfield and Murphy (2014) has some methodological limitations that could be potentially improved upon. Whilst the control group was tested later than the two experimental groups to allow them an equivalent amount of exposure to English through their timetabled lessons, this could present a confounding factor since the three groups were not all

exposed to an additional English lesson for 45 minutes per day for three weeks (which may have a novelty value of being out of normal class time) and the control group were older at posttest (although it is not clear how much older, which makes it difficult to judge the effect, if any, of this age difference and time delay). The data were analysed using the statistical method of MANCOVA, which has the assumption of independence, but given the repeated measures design this assumption was arguably violated since responses from the same learners across times constitute a dependency in the data. Alternative methods of data analysis (explored in section 4.5.2) that account for the hierarchical data structure, would potentially produce more robust results.

My intervention also builds on the work of Davis and Fan (2016), who designed a single group pre/post study investigating Chinese learners' (age 4–5 years) productive English vocabulary when presented with novel items either through singing or choral repetition. Davis and Fan (2016) found no statistically significant differences between vocabulary production of the five target items presented in either song or repeat conditions, compared to a control set of five words that were not presented, according to their mean length of utterance outcome measure. Arguably, having an active control condition would be more informative and ecologically valid than a 'placebo' or no-presentation control, since teachers are unlikely to be interested in a finding that suggests presenting vocabulary is better than not presenting it at all. It would be better to compare the experimental groups to a control that does what teachers usually do when presenting vocabulary. Furthermore, with only five items in each condition, the study is perhaps underpowered to detect any differential effects of songs or choral repetition, but it cannot claim (as the authors do) that there is evidence of no difference between the conditions either. As Davis and Fan (2016) is a single-group experimental design, it is challenging to draw reliable conclusions about the effect of the intervention on productive vocabulary. Also, given the repeated measures design, it would arguably be more robust to adopt a statistical analysis method that accounts for the hierarchical data, rather than using

ANOVA, which (like MANCOVA discussed above) assumes independence of the data points. It seems logical to extend this work to a between-groups design and test whether there is any difference in the relative effects of singing or choral repetition in comparison to actively teaching the target language in the control condition, whilst taking care to avoid presenting test items. This would provide a more robust test of the experimental method to check whether it is indeed testing that knowledge has been acquired during the intervention, since it should show that whilst all learners make progress, only those presented with the test items acquire these.

Both Campfield and Murphy (2014) and Davis and Fan (2016) focused on the aural and oral modalities, presenting, rehearsing and testing the FL through listening and speaking. This seems appropriate, given the young age of the participants and their lack of experience with the L2, as they are still developing their L1 reading and writing skills. There would potentially be effects of emergent L1 reading or writing skills on participants' ability to perform in an outcome measure that might confound a test of L2 knowledge. Maintaining the same modality through both presentation of materials and testing of knowledge is important to avoid a change in modality confounding the results (Murphy & Castillo, 2013). Focusing on listening and speaking then, Campfield and Murphy (2014) compared nursery rhymes with stories, whereas Davis and Fan (2016) compared songs with choral repetition. Choral repetition involved Davis and Fan's participants seeing a picture on the screen and being asked to repeat the associated phrase (e.g., seeing a picture of a lion and repeating *The lion is angry*) in unison. This choral repetition is not equivalent to Campfield and Murphy's prosodically salient nursery rhyme condition, which used traditional English nursery rhymes as the input. Table 3.15 summarises which input conditions from these two studies are melodic (i.e., have a tune) and/or prosodically salient.

Table 3.15. Summary of input conditions and melody/prosody

Input condition	Melodic	Prosodically salient
Song	Yes	Yes
Nursery rhyme	No	Yes
Choral repetition	No	No
Story	No	No

Another difference between these studies is that the input in Campfield and Murphy's (2014) study was equivalent across the groups, with the words of the nursery rhymes being reused to write the story. This controls for repetition, a common feature of songs with choruses and verses that follow a repetitive structure or repeat vocabulary items. There does not appear to be the same attention to making the input conditions equivalent in Davis and Fan's (2016) study. The songs they used are novel texts set to familiar melodies (e.g., *Twinkle Twinkle Little Star*) sourced from a textbook. The choral repetition input, however, is described as the teacher repeating the *phrase* (e.g., *The lion is angry*) twice per picture. It appears that the phrases are not equivalent to the songs, which are presumably longer and contain additional words and structures than the test phrases. In a third study comparing the effect of songs and stories on lexical acquisition (Albaladejo et al., 2018), the target words from the story (*balloon*, *kangaroo*, *snake*, *monkey* and *lion*) and story/song combination (*cake*, *strawberries*, *bananas*, *chocolate*, *fish*) are not matched for word length or number of encounters with the song (*doll*, *pill*, *bag*, *hat*, *bed*) condition. *Balloon*, for example, is encountered 17 times and *bed* three times. It is perhaps unsurprising that *balloon* is recalled more readily ($M = 0.52$, $SD = 0.51$) than *bed* ($M = 0.23$, $SD = 0.43$) in the posttest, and retained much more highly in the delayed posttest (*balloon*: $M = 0.59$, $SD = 0.50$; *bed*: $M = 0.11$, $SD = 0.33$). This lack of equivalence arises from the use of existing songs and stories, rather than creating new materials that would match across conditions like in Campfield and Murphy (2014). Even though it is ecologically valid to choose songs and existing stories that

teachers are likely to use in lessons, it compromises the internal validity of the experiment if the conditions are not equivalent.

To take this work forward, then, I decided to extend and build on the foundations laid by Campfield and Murphy (2014) and Davis and Fan (2016) by comparing songs (with melodies), chants (prosodically salient but no melody), a story condition (not prosodically salient), and an active control group that receive quality teaching in the target language but no presentation of the test items. The input materials would be designed to match across conditions, based on a set of traditional songs (since these are likely materials for FL teachers to draw upon) with the vocabulary reused in the story condition so that word length, token frequency and vocabulary word-band frequency are considered. Further considerations were to account for the hierarchical structure of data gathered through repeated-measures designs in my statistical analyses, and to ensure the study was suitably powered to detect any potential effects.

3.8.2 Warrant for the intervention study

My intervention study is based on comparing the relative effects of using songs, chants and stories to present learners with the FL with an active control group who receive quality FL lessons. This is well warranted because:

1. My systematic review indicated that presenting the FL through songs, chants or stories might be an effective way of enhancing young learners' acquisition of the FL (either specific vocabulary items, grammatical knowledge of word order, or spoken proficiency), but that methodological shortcomings of the existing literature require that this proposition is tested more robustly.

2. The comparison of songs with chants and stories is a meaningful one because teachers use these widely to teach FL. The addition of an active control group who also receive FL lessons means that the control group is not just a 'placebo' (Bishop & Thompson, 2024) but a more meaningful comparison with what teachers do when teaching the FL.

Making the control condition equivalent to the best currently available materials adheres to the values espoused by Wiliam (2018), who questioned the idea that research should be allowed to test new approaches that may or may not work whilst withholding approaches that we know do work. Instead of a 'placebo', therefore, I created control group conditions that reflect what teachers would have been teaching the participants at the same stage of FL teaching, using the best available resources suggested by the teachers.

3. There is little robust empirical evidence underpinning teachers' valued practice of using songs to help YLLs achieve specific linguistic outcomes. This intervention can therefore help to inform teaching practice by providing evidence of the relative effectiveness of songs, chants and stories, which may assist teachers' decision-making when planning activities to support their teaching.

3.8.3 Overview of the intervention study

In summary, drawing on the findings from my systematic review, and from my own and other teachers' experiences of using songs for teaching YLLs, Phase 2 of this project addresses whether there are measurable effects of using songs to achieve linguistic outcomes relative to other popular approaches, namely chants and stories. The design and methodology of the Phase 2 intervention and data analysis are the subject of discussion in the next chapter.

Chapter Four

Phase 2: Methodology

4.1 Aims and objectives of this research

4.1.1 Research questions

The research questions that drive Phase 2's intervention study are as follows:

RQ2: What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on:

a) all 22 stimuli?

b) a subset of fourteen previously encountered stimuli?

c) a subset of eight previously encountered stimuli containing novel vocabulary items?

d) their ability to correct a grammatical error in the *ne […] pas* negative word order in a subset of three previously encountered stimuli?

4.1.2 Hypotheses

The research questions give rise to four testable hypotheses respectively, as follows:

RQ2a –

H0: there is no interaction between input condition and time on participants' overall performance on the elicited imitation task.

H1: there is an interaction between input condition and time on participants' overall performance on the elicited imitation task.

RQ2b –

H0: there is no interaction between input condition and time on participants' performance on previously encountered stimuli in the elicited imitation task.

H1: there is an interaction between input condition and time on participants' performance on previously encountered stimuli in the elicited imitation task.

RQ2c –

H0: there is no interaction between input condition and time on participants' performance on stimuli containing novel vocabulary items in the elicited imitation task.

H1: there is an interaction between input condition and time on participants' performance on stimuli containing novel vocabulary items in the elicited imitation task.

RQ2d –

H0: there is no interaction between input condition and time on participants' ability to correct a grammatical error in the *ne […] pas* negative word order in the elicited imitation task.

H1: there is an interaction between input condition and time on participants' ability to correct a grammatical error in the *ne […] pas* negative word order in the elicited imitation task.

4.2 Experimental Design

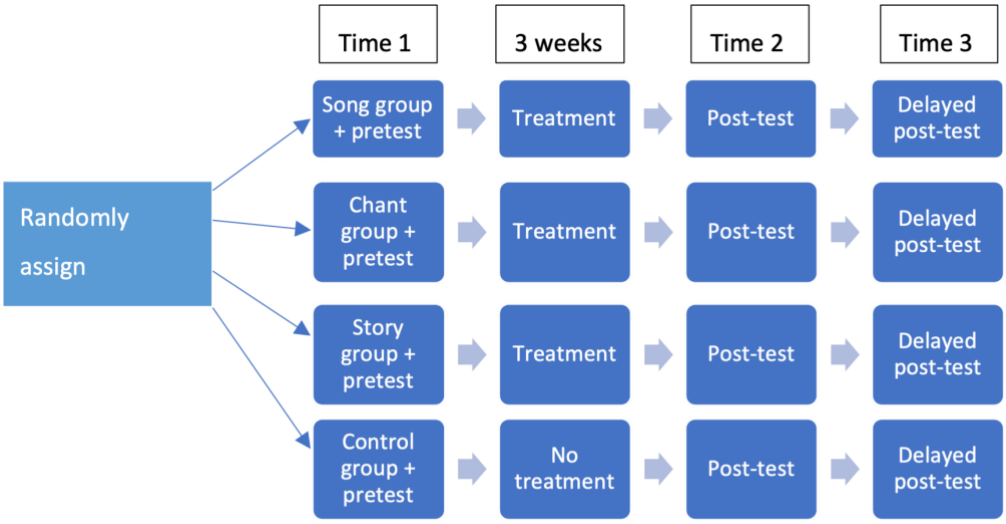
4.2.1 Trial design and rationale

This study was a randomised controlled trial to investigate whether any of the treatment conditions (song, chant or story) resulted in superior performance on the EIT compared to each other, and compared to the 'business-as-usual' experimental control. The independent variable was condition, and the dependent variable was the EIT outcome measure. Participants were randomly assigned to one of four conditions (song, chant, story, business as usual) in a 1:1:1:1 ratio with equal chance of being allocated to any condition. Performance on the EIT was measured at pretest, posttest and delayed posttest six weeks after the intervention ended. Figure 4.1 summarises the mixed experimental research design.

Using a randomised controlled trial design minimises threats to internal validity and is considered the most robust approach to evaluating causal relationships in experimental design (Bishop & Thompson, 2024; Campbell & Stanley, 1963; Connolly et al., 2017; Cook &

Campbell, 1979; Gorard, 2003, 2013; Shadish et al., 2002; Slavin, 1986). Random assignment to conditions means that participants have an equal chance of being allocated to any group, thus ensuring that average differences between those groups at baseline are the result of the play of chance, and not systematic differences, or biases. As a consequence, any differences at outcome can be more confidently attributed to the different interventions each group was exposed to.

Figure 4.1. Research design



Since I both conducted the intervention and assessed the outcomes, allocator-concealed and assessor-blinded procedures were used to minimise bias. The random allocation sequence was generated by an independent statistician with no knowledge of the nature of the experiment, to ensure that the allocation schedule could not be consciously or unconsciously subverted. Assessors were blinded to group allocation, to ensure that they would not be consciously or unconsciously influenced by knowledge of which group participants had been allocated to. Conducting the study across two (anonymised) schools based in south-west England increased

the diversity of the participant pool and enhanced the applicability of the findings to other primary school settings in England.

It was not possible to blind participants to their assigned treatment condition in this study since they would be aware they were singing, chanting, or reading a story. However, I avoided discussing details of the overall research design with the participants to minimise their awareness of the different groups as that might have influenced how they participated and thus their behaviour and outcome performance. They may have discussed differences between themselves outside of the lessons. Participants all knew they were having French lessons and helping me find out what works best for teaching children French. I did not tell them that groups had different materials. I made the input materials look as similar as possible in case they saw them during changeovers between groups. There is a potential contamination effect between conditions, which is a possible threat to validity, since participants in different conditions within the same school may have shared the intervention contents amongst themselves. I will address this potential limitation in the discussion (section 6.7).

The presence of the active control group provided a baseline for comparison with the treatment groups, thereby helping to isolate the effects of the treatment and permit attribution of differences in outcomes to the treatment conditions rather than to other factors. Without the active control condition, any potential null effects when comparing experimental conditions to each other would be difficult to disentangle from the possible failure of the outcome measure to detect any effects of treatment at all.

The repeated-measures design permitted assessment of any intervention effects over time, both immediately after the intervention and six weeks later. This provided greater insight into the effects of the interventions than would be possible with a single timepoint, cross-sectional design.

4.3 Context and participants

4.3.1 Context

The geographical context was a town in the south-west of England, in the UK. This covered an area of approximately 40.5 square kilometres (25 square miles) with an estimated population in 2021 of 132,416 (Office for National Statistics, 2024). According to the 2021 census (*anonymised* City Council, 2021), 22% of the local population identify as ethnic groups other than White British. 3 in 20 people were not born in the UK (14.4% of local population): 7.1% were born in mainland Europe, 4% in the Middle East and Asia, 2.1% in Africa, 1.1% in the Americas, and 0.2% in Oceania. English is the 'main' language of 92% of respondents. Of the 8% of respondents who did not choose English as their main language, Polish was the most frequently indicated language other than English (2%), followed by Romanian (0.7%), then Czech, Gujarati, Malayalam, Slovak, Bulgarian, Portuguese, Hungarian and Arabic (all <0.5%). French does not appear to be a frequently spoken language locally and can therefore be considered as a foreign language for the school population.

4.3.2 Sampling frame and sampling technique

This study's population was Year 3 (Y3, ages 7 and 8 years) pupils in state-maintained primary schools in England. Whilst random sampling from this population would produce the most robust and generalisable findings (Gorard, 2001), it was not logistically possible or practical for me to do this. I therefore used a mixture of purposive and convenience sampling based on the size (three-form entry), availability, and willingness to participate of schools within a 25km radius of my home.

4.3.2.1 Exclusion criteria

Children not attending formal educational settings (e.g., home-school groups) were excluded because the aim was to gather observations from a sample that can then be applied to mainstream educational settings. Atypical learners (for example children in special

educational needs [SEN] schools or spending most of their time in SEN units within their school) were not recruited, since any findings gathered from an atypically developing sample would not be applicable beyond the sample to the wider population of KS2 pupils. Children who are identified as having SEN within a mainstream classroom, however, were included as this is ecologically valid for mainstream settings where diverse pupil needs are commonplace.

4.3.2.2 Finding eligible schools

Schools were invited to participate based on several characteristics. Firstly, they needed to be within a reasonable travelling distance so that I could visit them to carry out the investigation and deliver the intervention – a process that would require daily visits over two to three months. Logistically, then, the schools needed to be within an hour's drive from my home.

Secondly, schools needed to have sufficient numbers of Y3 pupils to form four teaching groups. Some village schools have composite year groups because the number of pupils on the roll is very small. I therefore selected schools with three-form entry, meaning that approximately 80–90 Y3 pupils would be expected to be on the roll for September 2023, when the research was due to begin. This facilitated having large enough groups to run the intervention lessons, and also minimised how many schools were required to participate for practicality reasons as a solo investigator.

Thirdly, schools needed to begin formal French lessons in Y3, and not before. Y3 is the first year of KS2 (upper primary school), when MFL education becomes part of statutory provision for all pupils (DfE, 2013a). Given the wide variation in how and when schools meet the statutory provision targets (Holmes & Myles, 2019), focusing on Y3 beginner learners of French aimed to control for the variation in the extent and nature of MFL input that Y4–6 learners from different schools may have received. Schools where French is formally provided in Early Years and KS1 were excluded from participation.

4.3.2.3 Sample size

An *a priori* sample size to perform an appropriately powered *F*-test of difference was calculated using G*Power3 (Faul, Erdfelder, Lang & Buchner, 2007) for a 4*3 factor mixed ANOVA. Given the lack of reliable quantitative data in the systematic review findings (see Chapter 2), the effect size was estimated as .15, a conservative estimate of songs' effect on linguistic outcomes. Alpha was set at .05, and beta at .8, as recommended by Field (2018). Whilst this power calculation is analytic rather than simulation based and is likely to be too simplistic for a complex calculation of power for the mixed effects model I used for data analysis, it did nonetheless provide an estimate of how many participants to recruit given the scarcity of prior research with a comparable design in this field. Based on the power calculation, this study needed to recruit 108 participants.[14]

4.3.2.4 Recruiting schools

For reasons of practicality in administering the project and to minimise environmental factors influencing the findings by working with as few schools as possible, three-form entry primary schools within a 25km radius (n = 10) were invited to take part in the study by email to the headteachers in January 2023 (see Appendix B1a). A call for participants was simultaneously put out on social media (Twitter, Facebook, Instagram and LinkedIn; see Appendix B1b). Two headteachers immediately responded to the email and one further head teacher responded in April 2023, which was too late to participate. Two class teachers responded to the call for participants on social media. Initial email exchanges with the two class teachers resulted in both withdrawing due to the time commitment involved in the project. This left

[14] Following examiner feedback after my Confirmation of Status viva, I updated the G*Power analysis to comprise an ANOVA with fixed, special and main effects and interactions, with effect size 0.15 and df = 6 (4 groups*3 time points). For *alpha* 0.5 and *beta* 0.8, an estimated 612 participants would be required. Estimated power for a 0.15 effect size and n = 96 is *beta* = 0.15. This study is underpowered to detect a < 0.39 interaction effect size according to the updated G*Power calculation. Statistical power will be discussed in the Limitations section.

two interested headteachers from state-maintained primary schools, which are henceforth referred to as School 1 and School 2.

A telephone meeting was arranged to discuss the project with each school's lead MFL teachers to ascertain whether it was feasible to run the study in their setting. Both lead MFL teachers and their headteachers agreed to the timescale for the project to take place at the start of Year 3 in September 2023, before their pupils began any formal MFL lessons. They agreed to providing teaching space to run the intervention with four randomly-allocated groups but stated that the teaching may need to take place in two or three different rooms to accommodate existing class timetables, extra-curricular activities, and pupils' learning support requirements. Both schools wished to run the intervention classes without interrupting their morning maths and literacy lessons, which meant that I needed to stagger the baseline data collection and intervention periods as I could not timetable teaching in one school in the morning, and one in the afternoon as I had imagined doing. It was thus decided that School 1, who also agreed to give me access to their current Year 3 group to run the pilot study in June 2023, would be first to embark on the project in September 2023, and School 2 would begin their baselines in October 2023. Furthermore, the liaison teacher in School 1 would be on maternity leave from October half term, thus it made logical sense to work together before her leave began. A timetable was drawn up (Appendix B2) to reflect the projected amount of time required for baselines, intervention and posttest data collection periods in each school once I knew how many children would be participating from each school.

School 1 had projected 84 and School 2 had projected 90 children on the roll for the 2023/24 cohort of Y3 pupils at the time discussions began in March 2023. This provided a potential eligible sample of 174 participants.

4.3.2.5 Characteristics of participating schools

This section describes the characteristics of the two participating schools, with information drawn from their respective websites or Ofsted report, or school information provided on the local council website, unless stated otherwise. References to the schools' websites, name of local council, and Ofsted reports are not cited to preserve participant anonymity.

School 1

School 1 is a state-maintained community primary school located in a culturally and linguistically diverse urban neighbourhood. 11.4% of the local population were born outside of the UK, compared to 14.4% of the overall local population (ONS, 2024). There were 525 pupils on the roll as of January 2023. It was rated 'Requires improvement' by Ofsted in 2021. According to 2022/23 data, the school population contains 22.9% pupils eligible for free school meals, just under the national average of 25.9%. The school has 14.1% of pupils with SEN support (higher than the national average of 13.5%) and 19% with EAL status (just below the national average of 22%). Y3 and Y4 are taught in six composite classes rather than in two separate year groups. The Y3/4 curriculum introduces French greetings, numbers and colours, and feelings. The school has a specialist MFL lead teacher, who at the time of this study taught in Y5, and a Y3/4 teacher who was being trained to cover the MFL lead's maternity leave from November 2023. There is a team of teaching assistants to support behaviour and learning across the Y3/4 cohort.

School 2

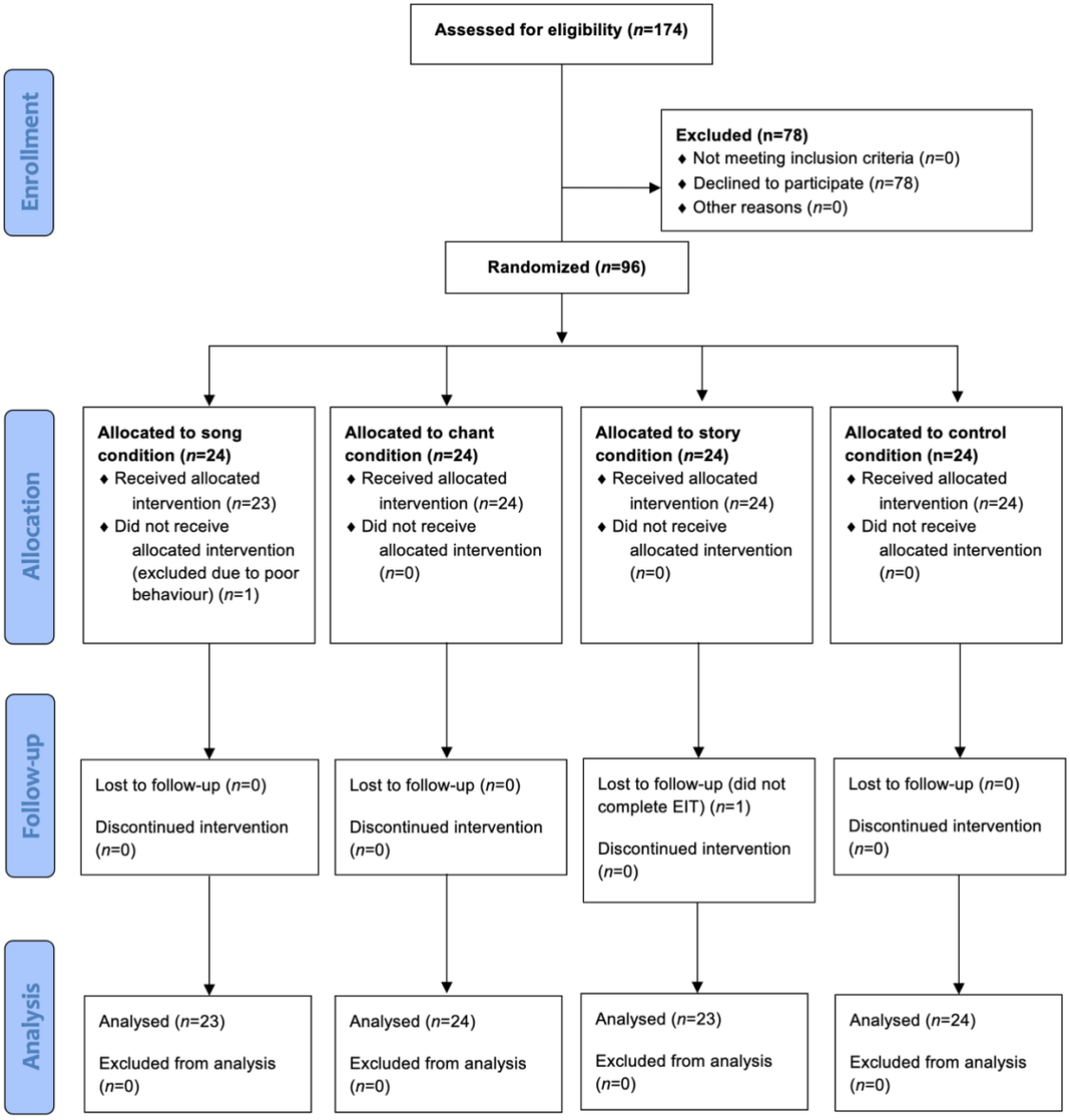
School 2 is a state-maintained community junior school located in a suburban area where 2.8% of the population in the local school area were born outside of the UK, compared to 14.4% of the overall local population (ONS, 2024). There were 361 pupils on the school roll as of January 2023. It was rated 'Good' in its most recent Ofsted report in 2024. According to 2022/23 data, 12.7% of pupils were eligible for free school meals. 6.6% of pupils had EAL

status. 15.2% of pupils received SEN support, higher than the national average. Y3 consisted of three class groups. There were two MFL specialists in the Y3 teaching team, and two teaching partners (for behaviour and learning support). The Y3 curriculum for French included topics such as "I'm learning French", animals, fruit, musical instruments, and phonics instruction.

4.3.3 Recruiting participants

After receiving approval from participating headteachers and class teachers at both schools, the information sheet (Appendix B3) and consent form (Appendix B4) for parents and children was distributed by the schools to their Y3 pupils in early September 2023. The liaison teachers in each school collected the consent forms and returned them to me. Parents received reminders from the school to return consent forms at their earliest opportunity. At the end of the recruitment window, parents of 29 pupils in School 1 and 67 pupils in School 2 had returned signed consent forms giving permission for their children to participate in the study. There was a much higher proportion of consent forms returned in School 2 than School 1, perhaps due to the schools' different demographics and levels of parental engagement. There were no dropouts. 96 participants thus participated in the trial, as visualised in the flow diagram in Figure 4.2.

Figure 4.2. CONSORT flow diagram[15]



4.3.3.1 Participants

96 participants were recruited, 29 children from School 1 and 67 from School 2. Given the power calculation recommended recruiting 108 participants, this study is likely to be underpowered. Initial sample characteristics are shown in Table 4.1. Gender and age were determined by asking participants for their gender and age in years and months. All

15. CONSORT ('Consolidated Standards of Reporting Trials') offers guidance and materials, including the flow diagram, as a "standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation." (CONSORT, no date).

participants were in Y3, with a mean age of 91.92 months (*SD* = 3.67). I administered a language background and music questionnaire (see 4.4.3.3) to gather information about children's L1 background and musical experience. 24 participants reported having exposure to a language other than English at home, including Romanian, Polish, Turkish, Lithuanian, Welsh, Arabic, Hungarian, Cantonese, Tamil, Filipino, Gujarati, Hindi, Czech and French Creole. I spoke to the French Creole speaker in French but she was unable to respond and said she did not understand me. One boy reported speaking Polish and only watching Polish TV at home, but upon corroborating this with his teacher, he was found to be a monolingual English speaker (see 4.3.3.2). He is not therefore included in the number of participants with EAL status. 22 participants had received music lessons with 15 continuing to receive weekly instrumental lessons at the time of data collection. Five participants had been playing an instrument for more than two years and reported practising for more than 30 minutes per week.

Table 4.1. Characteristics of the initial sample

School	Male	Female	Mean age (months)	SD age (months)	EAL status	Total
1	14	15	90.07	3.92	11	29
2	27	40	92.72	3.27	12	67
Total	**41**	**55**	**91.92**	**3.67**	**23**	**96**

4.3.3.2 Language background and music questionnaire procedure

To establish children's L1 status and exposure to languages other than English outside of school, and their level of experience with musical instrument instruction, a short questionnaire (Appendix B5) was administered orally. The three questionnaire sections explored children's L1 interactions with their caregivers, exposure to L1 through different media, and musical instrument instruction and practice time through a multiple-choice answer format. I explained the purpose of the questionnaire and discussed children's answers with them to help them

understand the questions. For example, one child confused their parent having an Irish accent with them speaking Irish, which needed clarification. I recorded participants' answers onto a spreadsheet. Each section received a score, and a higher score reflected more exposure to an L1 other than English, or more time spent learning musical instruments. These data enabled comparison of outcomes to take prior musical experience and L1 exposure into account.

Since the questionnaire involved participants' self-report, it is unlikely to be externally valid. Indeed, as mentioned in section 4.3.3.1, one participant reported mainly speaking Polish and being extensively exposed to Polish media at home, but I did not feel his answers were accurate as he seemed to be exaggerating. When I corroborated his EAL status with his teacher, he apparently did not have Polish parents and his family (who the teacher knew well) had no connections with Polish. This demonstrates the fallibility of such a measure. However, the questionnaire also permitted more nuanced understanding of whether children who reported having EAL status mainly heard or spoke their L1 at home, or mainly heard or spoke English. The administration of the questionnaire also helped to build a rapport with the participants at the start of the baseline testing period.

4.3.3.4 Screening variables

In addition to the language background and music questionnaire, the following measures were administered (reported as *construct: test administered*):

1) Non-verbal IQ: WASI FSIQ-2 measure of non-verbal reasoning (Pearson, 1999)

2) English receptive vocabulary: English Picture Vocabulary Size Test (PVST; Anthony & Nation, 2021)

3) French receptive vocabulary: Échelle de vocabulaire en images Peabody (EVIP; Dunn, Dunn & Thériaut-Whalen, 1993)

4) Rhythm: Children's rhythm synchronisation task (c-RST; Ireland, Parker, Foster, & Penhune, 2018)

These measures are age appropriate and similar to normal classroom activities.

Non-verbal IQ (WASI FSIQ-2 measure of non-verbal reasoning)

The Weschler Abbreviated Scale of Intelligence (WASI; Pearson, 1999) contains four subtests – Vocabulary, Block Design, Similarities, and Matrix Reasoning – which tap into different facets of intelligence including verbal knowledge, spatial and non-verbal reasoning, visual information processing, and crystallised and fluid intelligence. The four subtests comprise the full scale, called FSIQ–4, which takes 30 minutes to administer but the FSIQ–2 is sufficient to provide a general summary of individual participants' cognitive functioning. In coordination with the PVST measure of English receptive vocabulary, described below, the Matrix Reasoning test was deemed sufficient to establish a baseline cognitive measure of pupil participants in this study since a diagnostic level of testing is not required. The Matrix Reasoning (a non-verbal measure) includes maximum 24 items and takes up to 10 minutes to administer per participant. The WASI is suitable for use with a broad age range of 6–89 years. It can be administered by doctoral students who receive appropriate training and support for interpreting the results from their supervisor, as is the case in this study. The Matrix Reasoning raw scores are converted to a *t*-score to produce a standardised test result that takes into account the different ages of participants to permit comparison with each other.

The reliability of the WASI subtests were estimated from a single administration of the test split by rank order of difficulty in two halves, with the variations of total scores for each half found to not be statistically different from each other. The internal consistency reliability coefficient of the Matrix Reasoning test for the 7–8-year age group is .94, above the average of .92 for the children's sample. The test-retest reliability of the children's sample was .76, based on 61 children aged 6–11 taking the test on two occasions (average 31 days apart). These coefficients indicate stable and consistent scores for the Matrix reasoning test in the 7–8-year age group within itself and across time.

Regarding the Matrix Reasoning validity, a multipronged and long-term process has established its content validity to ensure the subtests provide a good estimate of *g*, a person's

general intellectual ability. Content analysis has established that WASI subtests (i.e., Matrix Reasoning, Vocabulary) share similar parallel content to the full Weschler batteries (WISC-III and WAIS-III). The WISC-III does not have a Matrix reasoning test, but the correlation with the WAIS-III Matrix reasoning (.66) indicates that the WASI FSIQ-2 subtest measures similar constructs to the full-scale version. The construct validity, namely whether the WASI FSIQ-2 does indeed measure the verbal and non-verbal intellectual traits of interest, has been established both by Pearson (1999) through statistically significant moderate intercorrelations of the subtests, correlations with IQ scales and confirmatory factor analyses, and independently by comparing the WASI-II with the RIAS-2 (Reynolds Intellectual Assessment Scales; Reynolds & Kamphaus, 2015; Sopoci, 2023). All analyses indicate that the WASI FSIQ-2 is a reliable and valid measure of general intellectual functioning, and that the Matrix Reasoning is a relatively accurate measure of non-verbal intelligence that compares well to more in-depth IQ scales.

English receptive vocabulary knowledge (PVST)

This test of English receptive vocabulary was developed by Anthony and Nation, with the version used in this study released in 2021. It is an online test that uses picture choices with oral and written cues and was designed for use with pre-literate or literate native or non-native speakers of English. The test was adapted for use on Gorilla software by Schulz (2024) for use in a similar doctoral study with primary school learners of German as an MFL.

The PVST begins with a training phase of five items where participants learn to listen to the audio prompt, then select the most appropriate of four multiple-choice pictures to match the oral prompt. After the training phase, the test phase of 96 items is divided into two sections with a motivating message after 48 items. The PVST produces an estimate of participants' vocabulary size in terms of bands of frequency of thousands of words known, which ranges from 1000 to 6000 most frequent word families for young native-speakers of

English. Each test word represents 62.5 words in the source list. The participants' raw scores are multiplied by 62.5 to indicate their total vocabulary size, a process that happens automatically as part of the program's calculated results. A child with a score of 54 thus has an indicated receptive vocabulary size of 3,375 word-families. The PVST has undergone several rounds of improvements to increase its validity (Anthony & Nation, 2021). Although there is scant published evidence of the PVST's reliability to date, with trials only including two children and a group of adults (Anthony & Nation, 2021), the multiple-choice picture vocabulary test format has a long history in assessing young learners' receptive vocabulary knowledge (e.g., Dunn & Dunn, 1981) and Nation is a well-cited expert in measuring children's vocabulary size (e.g., Bauer & Nation, 1993; Nation & Anthony, 2016). The test was also used successfully in Schulz's (2024) doctoral thesis with learners of German (age 7) in the UK primary school context.

French vocabulary knowledge (EVIP)

Whilst participants in this study were beginner French learners, taking a baseline measure of their French vocabulary knowledge established their beginner status more objectively and with more nuance as a continuous variable than taking their teachers' word for their beginner status as a binary variable (beginner/not beginner). It could also help account for any variance in the outcomes predicted by prior French knowledge by controlling for French vocabulary knowledge in the baseline covariates.

The chosen French vocabulary measure, the Échelle de vocabulaire en images (EVIP) French vocabulary test (Dunn et al., 1993), is the equivalent of the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1981), a receptive vocabulary measure used in several studies in this field (Albaladejo et al., 2018; Leśniewska & Pichette, 2014; Schunk, 1999). The test-retest reliability of the EVIP is r = .78 for age 4–10 years (Dunn et al., 1993). It is slightly lower at r = .69 for age 7;0–7;11 and r = .74 for age 8;0 to 8;11, the age groups of

my participants, but this still represents high test-retest reliability. In terms of content validity of the measure, the EVIP vocabulary covers eighteen topics familiar to children from any francophone background (not just Canada, where it was developed). The vocabulary lists match those of the PPVT (Dunn & Dunn, 1981), which have been used for 30 years and accepted by the field as a representative measure of receptive vocabulary knowledge that differentiates higher and lower ability across the age range, and no gender bias (Dunn et al., 1993). Whilst no external validity research was available for the EVIP itself, the PPVT on which it is based correlates highly (r = .71) with other vocabulary measures (Dunn et al., 1993). Overall, then, the EVIP appears to be a reliable and valid test of children's receptive French vocabulary knowledge.

The researcher presents a page of the test booklet with four black and white images displayed, naming one of the pictures orally (e.g. "Can you point to *le chien* for me?"). The child points to the picture that best corresponds to the target word. The test takes 5–10 minutes to administer to each participant. There is a practice phase with three items, and then a maximum of 45 items in the test. After eight or more consecutive errors, the ceiling level is considered to have been reached and the test finishes.

Children's Rhythm Synchronisation Task (c-RST)

To account for children's aptitude for rhythm, which may have a confounding effect on their ability to learn through prosodically salient input, I administered a rhythmic ability test. The c-RST is a computer-based production task developed by Ireland et al. (2018) to assess children's ability to tap in time to a series of rhythms of differing metrical complexity. The c-RST is adapted for children from musical ability tasks developed for adults (Chen, Penhune & Zatorre, 2008). The children's version is easier than the adult task (i.e., the highest level of metrical difficulty was removed), shorter to administer, and includes simple graphics and an engaging storyline (about a giraffe tapping rocks to extract the gold inside them). There are

three levels of metrical difficulty: the lowest level is strongly metric, thus easiest to follow along with; the medium level is less strongly metric; and the hardest level is weakly metric, hence more unpredictable and harder to tap along to (see visualisation, Figure 4.3).

Children are first asked to listen to a tapping rhythm on the laptop screen whilst the giraffe's headphones are illuminated, and then to tap along on the mouse or keypad whilst the giraffe's hoof is illuminated (see screenshot, Figure 4.4). Six rhythms, two per level, are presented in counterbalanced order for three consecutive trials (18 total). There is a practice phase of five strongly metric trials before the test phase, during which the experimenter helps the participant learn how to complete the task and gives feedback. These five rhythms are not repeated in the test phase.

Scoring. The outcomes measured are (1) percent correct, defined as the child's ability to tap within the 'scoring window' (half the interval preceding/following the stimulus tap), and (2) percent inter-tap interval (ITI) synchrony, which measures the child's ability to reproduce the same temporal structure of a rhythm. The ITI synchrony measure is taken from the ratio of the child's response intervals (r) to the stimulus time intervals (t) as follows: Score = 1–abs(r–t)/t. The values obtained for measure 1 and 2 are multiplied by 100 to produce a percentage score for correctness and ITI synchrony.

The c-RST provides standardised scores for age groups from 7–13 years, for children who have prior musical training (operationalised as children with 2.5 consecutive years of extra-curricular 30-minute weekly music lessons plus 30 minutes practice per week) and children without equivalent music training (i.e., less than 2.5 years' similar experience). The c-RST therefore accounts for children who may have developed advanced rhythmic ability through musical training, and age-related increased rhythmic ability.

As a baseline test that does not differ for musicians/non-musicians, the third part of the c-RST is a tapping and continuation task (TCT) testing basic synchronisation and timing ability. In the TCT, a regular metronome rhythm plays for fifteen seconds for six trials of

identical tempo. Children tap along with the metronome, then continue the same beat for a further fifteen seconds once the metronome stops. The variability of the child's tapping is measured across the six trials, scored separately for the paced/non-paced trials, with the standard deviation of the inter-tap interval (ITI) for the six trials averaged across paced/non-paced tapping. A coefficient of variation is calculated by dividing the average SD by the average ITI, in other words a score of each child's tapping variability relative to their own performance. The TCT serves as an auditory motor and cognitive control task for the c-RST (Ireland et al., 2018).

Figure 4.3. (taken from Ireland et al., 2018). Examples of rhythm stimuli from c-RST: strongly, medium, and weakly metric (in order of regularity of rhythmic pulse, where strong = easiest). Figure adapted from Tryfon et al. (2017)

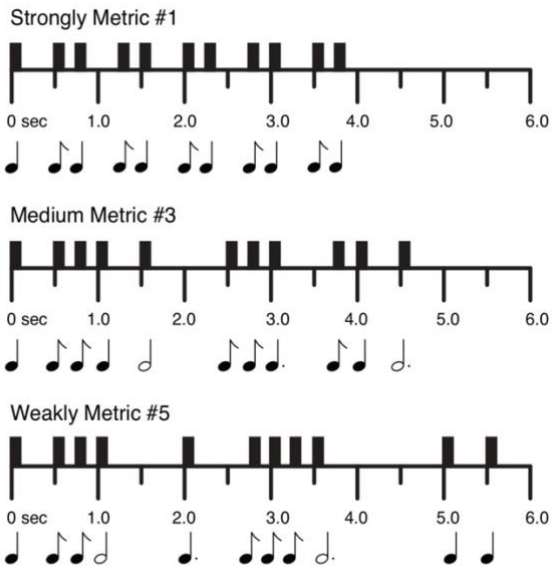
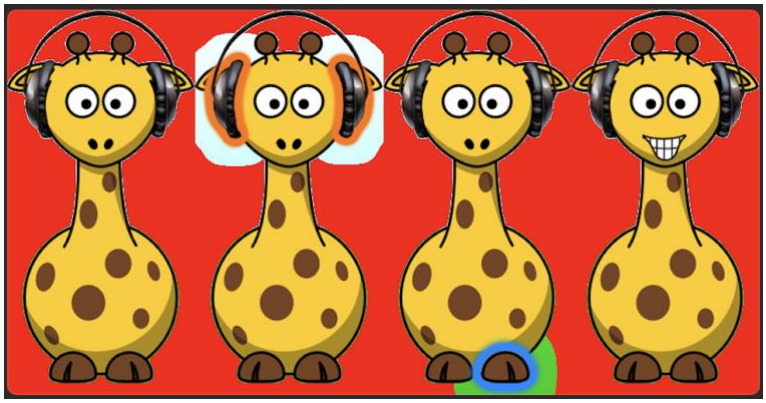


Figure 4.4. Screenshot of graphical display from c-RST task (presented one image at a time in the task) showing the giraffe with headphones/hoof that animate for part 1 and 2 of the task



Internal-consistency reliability of the c-RST was examined using Cronbach's alpha, estimating the mean of all possible split-half reliabilities (Ireland et al., 2018). The test was found to be adequately reliable for musicians ($\alpha = 0.64$) and marginally less so for non-musicians ($\alpha = 0.60$). The authors claim to have developed a more valid test of children's emerging rhythmic skills, with musical ability divided into rhythm synchronisation and melody discrimination which develop at different ages, and assessed independently of maturation, cognitive and motor abilities, and taking into account prior musical training (Ireland et al., 2018).

It should be noted that I chose not to test the participants' melody discrimination using Ireland et al.'s (2018) materials, since the screening testing burden on the young participants was already substantial. The potential limitations of this choice are discussed in section 6.7.

4.3.3.5 Administration of screening variables

Screening variables were administered in two ways. For all 96 initial participants, c-RST and EVIP were administered one-on-one at a desk in a quiet room with the EIT pretest. Testing session length varied according to each child's attentiveness, response time, and comprehension but did not take more than 30 minutes per child. Tests were administered in a

random order for each child to minimise any order effects. For School 1, the PVST and WASI measures were administered in small groups in the provided teaching space (a break-out room situated along the corridor from the Year 3 classrooms). In School 2, the PVST and WASI measures were administered in classrooms with whole class groups.

c-RST

The rhythm synchronisation task was presented on a Windows Surface laptop, model i5, running the NeuroBehavioural Systems Presentation software package (version 23.1). The laptop had a mouse trackpad but a plug-in mouse was used to help the children with the motor skills required for clicking (as explained in Appendix B6.3 on piloting). It took 15–20 minutes to administer the c-RST per child. The participants enjoyed the c-RST much more than the pilot participants had, which could be attributed to them being a year younger and thus more interested in the game's outcome.

EVIP

French vocabulary knowledge was tested on a one-to-one basis using the EVIP booklet and spreadsheet to record answers, as described in 4.3.3.4.

PVST

The test of English receptive vocabulary was administered using school laptops (School 1) or iPads (School 2). In School 1, given the limited seating capacity of the room available for completing the screening variables, the test was administered to participants in small groups of 2–3 pupils. Participants were given a set of headphones to listen to the audio after hearing the instructions on how to complete the exercise. Some then required further assistance during the practice phase of five items to get used to the controls for tapping 'next' or selecting their chosen picture. Generally, the test ran smoothly after this practice phase and took 10–15 minutes to complete. In School 2, where a class set of iPads and headphones was available and most pupils were engaged in the research, the test was administered to a whole class

group at once, on three consecutive afternoon sessions. One participant wore hearing aids and thus completed the PVST on her own the next morning in the quiet room with her other screening variables. Only one participant failed to complete the test on their first attempt as they kept exiting and restarting the test by accident. They completed the test on the next morning one-to-one with the researcher in a quiet spot outside the classroom where they usually received learning support each morning.

WASI

The Matrix test was administered on paper with answers subsequently recorded on an Excel spreadsheet, as described in the pilot study, Appendix B6.3.

Language background questionnaire

The language background and music questionnaire was administered one-to-one with each participant at the first baseline meeting, with answers recorded on a spreadsheet. There is some possibility of children reporting answers that they think researchers want to hear when they engage with us for the first time, as evidenced by the participant who falsely claimed to speak Polish at home (see 4.3.3.2). There was also some confusion about whether having a different accent (e.g., an Irish accent) constituted a parent speaking a different language. This confusion was mitigated by asking class teachers to confirm that pupils with an L1 background other than English were indeed EAL pupils, in the cases where I was unable to ascertain this from the pupil themselves beyond reasonable doubt.

4.3.4 Group characteristics

Group sizes and participant characteristics after random allocation to conditions are shown in Table 4.2.

Table 4.2. Participant characteristics after random allocation to conditions

Group	Male	Female	Mean age (months)	SD age (months)	EAL status	Total
Song	8	15	91.81	3.66	7	24
Chant	8	16	91.99	3.67	9	24
Story	12	11	91.97	3.63	3	24
Control	12	12	91.92	3.65	5	24
Total	**40**	**55**	**92.02**	**3.64**	**23**	**96**

Figures 4.5 to 4.9 visualise the distributions of the screening variable scores across groups. As anticipated through the use of random allocation to conditions, groups appear to be homogenous on visual inspection of these violin plots, although some descriptive variation is evident. Table 4.3 summarises the mean and standard deviation for each group on the screening variable measures. It would not be appropriate to conduct statistical significance tests to determine homogeneity of the groups, since nothing can be concluded from a null result (i.e., where $p > .05$). Hence the descriptive statistics are sufficient to ascertain group similarity.

Table 4.3. Descriptive summary of screening variables per experimental group

		WASI matrix		English vocabulary		French vocabulary		c-RST % correct		ITI % deviation	
Group	**Count**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Song	24	49.38	6.51	4369.79	820.49	10.21	7.42	77.64	8.02	44.18	9.79
Chant	24	48.29	8.61	4273.44	932.83	12.29	6.79	75.82	9.31	43.29	11.88
Story	24	47.67	9.11	4406.25	869.16	9.71	4.61	73.56	17.80	40.21	15.13
Control	24	49.17	6.64	4455.73	912.71	11.79	7.68	72.82	17.64	42.08	13.29

Figure 4.5. Non-verbal IQ scores group comparison

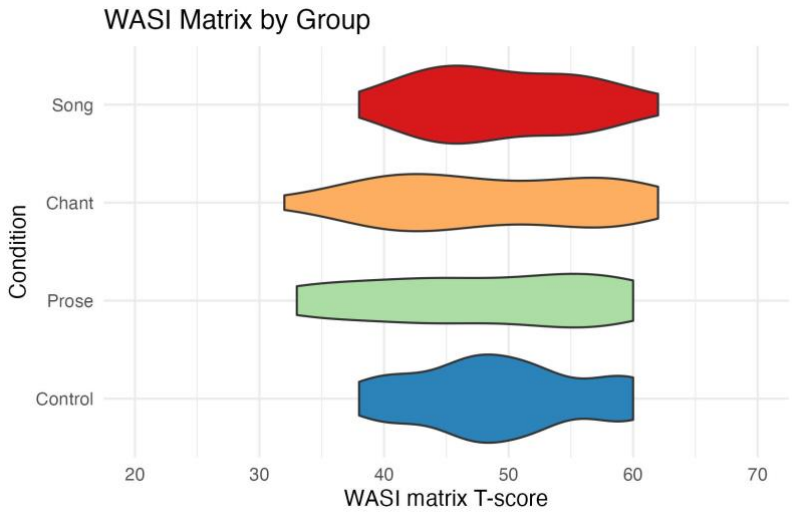


Figure 4.6. English vocabulary scores group comparison

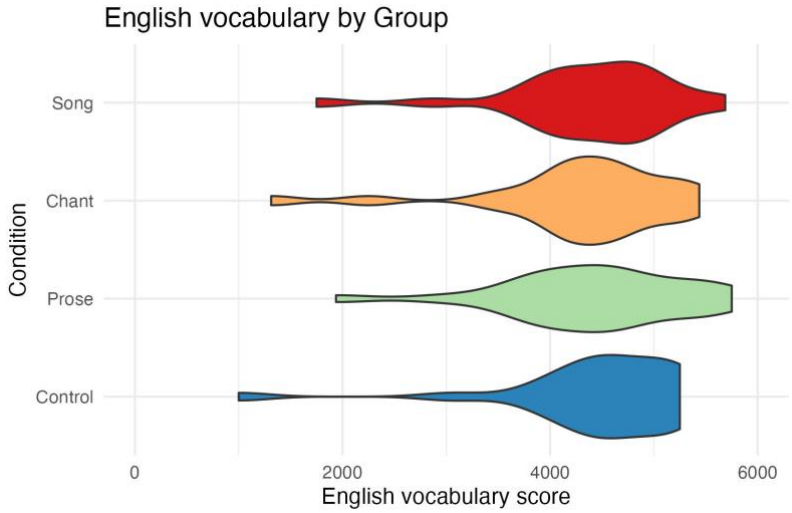


Figure 4.7. French vocabulary scores group comparison

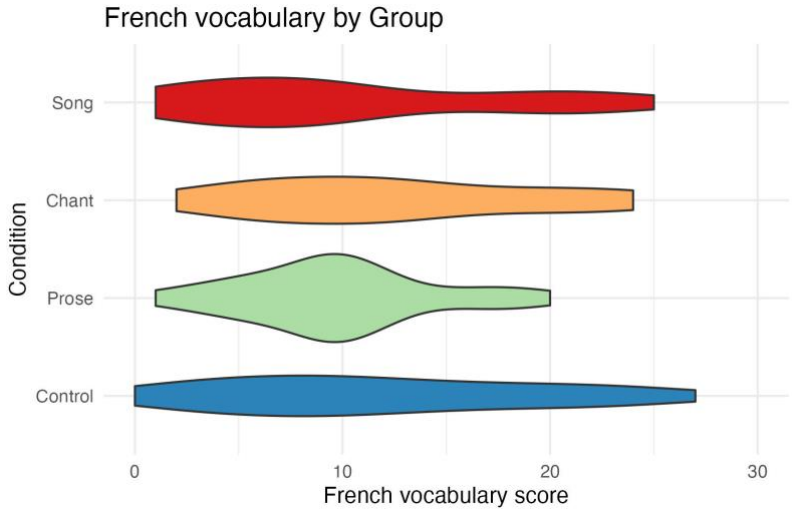


Figure 4.8. % Correct taps on rhythm synchronisation task scores group comparison

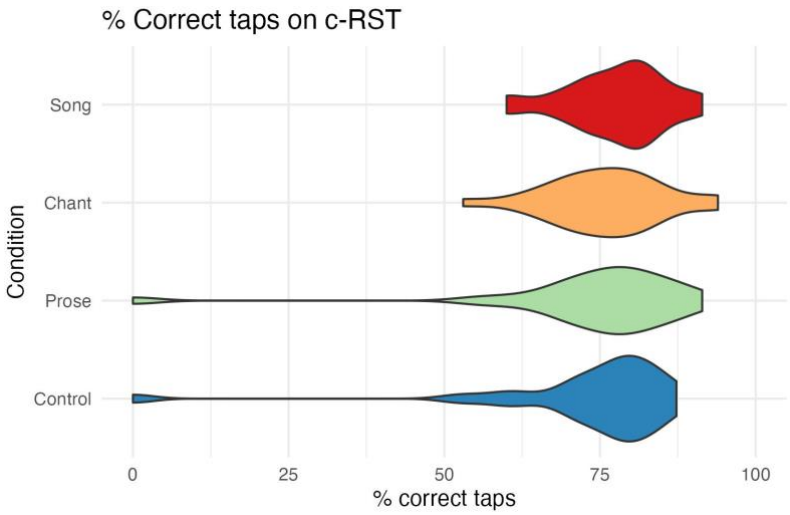
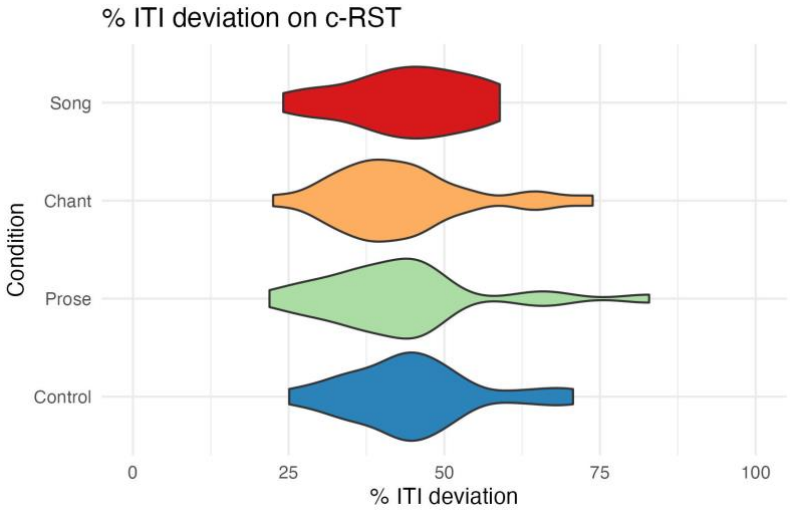


Figure 4.9. % Inter-tap interval deviation scores group comparison



4.3.5 Attrition and final sample size

Two male participants from School 1 dropped out of the study. The first completed the baseline tests and one intervention lesson before being removed from the group due to his unmanageably poor behaviour that distracted all other participants in his group and prevented the intervention being taught as planned. On discussion with his teacher, I discovered that he usually had one-to-one behaviour support in lessons, which was not feasible for the intervention group as his teaching assistant was needed in class. Given that the intervention needed to proceed as planned and with fidelity to condition, which is already challenging in a

real-world context in schools, this meant that he could not continue to participate in the study. As a teacher, this saddened me as I wished to teach all of the children who had consent to take part. However, my position as a researcher with a very limited amount of time to gather the data and a desire to maintain the highest standards of data collection possible was at odds with my position as a teacher in this aspect of the project. I continued to interact with him when I saw him around the school and gave him a certificate of participation at the end of the project with the other participants. A second participant declined to complete the outcome measure assessments at any of the three timepoints after the assessment began, thus in spite of willingly participating in the intervention lessons and completing the screening variables, he was also removed from the final data analysis. With this attrition (n = 2) the final sample size was 94, thus further slightly reducing the power of the statistical analyses. Final sample participant characteristics are shown in Table 4.4. These two drop-outs did not change the comparability of the groups.

Table 4.4. Characteristics of the participants (values in brackets show original sample statistics)

School	Male	Female	Mean age (months)	SD age (months)	Total
1	12 (14)	15	90.30 (90.07)	3.97 (3.92)	27 (29)
2	27	40	92.72	3.27	67
Total	**39**	**55**	**92.02** (91.92)	**3.64** (3.67)	**94** (96)

4.4 Interventions

4.4.1 Overview

This trial was designed to assess the relative effects on children's French proficiency of being exposed to French input through songs, chants or stories, compared to each other and to an experimental control condition. This section describes how the materials were created for the intervention and how they were administered. The materials design process was informed by procedures outlined in Campfield (2010) and Campfield and Murphy (2013, 2014), which as

described in section 3.8, provided a starting point for this study. My intervention added a song condition to the rhythmically-salient and prose conditions in Campfield and Murphy's earlier work. The aim was to build on the foundations laid by Campfield and Murphy, and see if their findings could be replicated in a new context with similar aged participants.

4.4.2 Procedure

All groups received 242 minutes of French input delivered by the researcher (an experienced French teacher). There were 22 minutes per lesson on 11 days across three weeks (four lessons in week 1 and 2, and three lessons in week 3 – see lesson plans in Appendix B11c). The intervention took place during the school day, as part of the normal timetabled lessons, not as an after-school or lunchtime activity. Participants were withdrawn from class to take part in the intervention, which replaced the usual MFL input that they would have been having at the start of Y3. Each intervention lesson lasted for 22 minutes. I created a schedule for the intervention lessons that I shared with the Y3 class teachers, with each group rotating the order in which they had their lessons. On the first day, the song group went first, followed by chant, story and control groups. The second day, the chant group went first, followed by story, control and song groups. This order rotation continued, to minimise any effects of groups being disadvantaged by having all of their sessions before lunch or last thing in the afternoon when participants are potentially hungry or tired.

4.4.3 Materials

4.4.3.1 Song and chant materials

24 songs were selected from books, CDs and online materials I have collated during my teaching career (see Appendix B11a). The first criterion for selection was that I enjoyed the song and felt comfortable singing it for the recorded materials, since any aversion or reluctance on my part to sing the song might influence how the participants perceived it. Secondly, since many French *comptines* (children's songs) are chanted with minimal melody,

this limited the choice of songs to those with a strong melody to differentiate them from the chant condition materials. In the chant, the prosodic structure takes precedence over the rising and falling pitch contours of the melody. The songs contained the same prosodic structure as the chants, with the additional pitch contours of the melodies. In all other respects, the song and chant conditions were identical. Two songs or chants (or three if they are very short) were presented per lesson. Nine songs or chants were thus presented in weeks 1 and 2, and six in week 3. At the start of each session from the second session onwards, the previous day's materials were sung or chanted once through. Thus all materials were met twice except for the final day's materials. Some children immediately enjoyed the early songs and requested them every time, which as a teacher was challenging to deny. Nonetheless, for the sake of fidelity to treatment, I explained that we could not sing/chant the materials more than we did during the experiment, but that I would give the schools all of the materials for them to enjoy after we had finished.

4.4.3.2 Story materials

The story condition consisted of 11 instalments of a story that I created for the purposes of this study (see Appendix B11b). One story instalment was approximately the same length as two songs. One instalment was presented per lesson, except on the second day when there were two shorter instalments. Care was taken to make the story engaging and follow a classic narrative structure that was likely to be familiar to the participants. The story followed the *lapin*, a rabbit who was feeling sad, and his friend the *hibou* (owl) through their adventure in the forest. *Hibou* introduced *lapin* to other characters from the songs such as Frère Jacques, the mill owner, and the elephants and other animals, in an attempt to cheer him up. Nothing succeeded until *lapin* saw the elephants dancing on a spider's web at the end of the journey. The children were highly engaged throughout because of the possibility that *lapin* was correct in suspecting that the *loup* (wolf) was coming to eat him. Happily, *lapin* turned out to be wrong and everyone lived happily ever after, much to the children's relief. This narrative arc

involved the introduction of the characters and the 'problem', the characters' attempts to solve the problem, and finally the resolution.

4.4.3.3 Control materials

The 'business as usual' experimental control group received the schools' planned French lessons, which I delivered myself to minimise the possibility of teacher effect confounding the result. To differentiate them from the experimental groups' materials, the control materials took a more explicit, decontextualised, focus-on-forms approach, with single vocabulary items presented in thematic groups (e.g., colours, numbers, greetings) rather than rich and complex input comprising longer passages or whole sentences. This content was determined in discussion with the participating schools, who both used the Primary Languages Network materials. It was decided to begin with the Y3 introductory materials as would have been the case at that stage of the year. I adapted these to create 11 mini 'units' of work with greetings, introductions, numbers, colours, animals, days of the week and months of the year as the themes. Since songs are popular resources with KS2 MFL teachers (Hamilton & Murphy, 2023), they would likely have normally formed part of the schemes of work for this age group. This presented a confounding factor because the effect of songs in the experimental condition could not be isolated as a variable if songs were present in the control condition. A balance needed to be struck between presenting the control group with an ecologically valid comparison (which removing all songs for three weeks might have compromised), and not confounding this study's results. After some deliberation I decided that the control materials should not contain songs or that would confound the study findings. This would have been the case even if the songs were not traditional French songs, since some MFL songs are especially written for explicit teaching of forms and their prosody is less 'naturalistic' compared to folk songs composed in French. Instead, I created engaging materials with illustrations that matched the three experimental conditions in design, and delivered the

materials in a lively, energetic fashion to which the children responded just as enthusiastically as the other groups.

4.4.3.4 Presentation of input materials

Experimental conditions

Input in all three experimental conditions was presented through a PowerPoint on a large screen (either an interactive whiteboard or projector screen, as available in each classroom). Text was accompanied with visually stimulating illustrations to aide comprehension, and pre-recorded audio to ensure that delivery was true to condition (i.e., no accidental melody in the rhyme condition, which might be difficult to maintain if teaching 'live' since the materials are otherwise identical). Figure 4.15 provides an exemplar screenshot of the PowerPoint presentations of all four groups' materials. The only difference between song and chant conditions was the audio file (with or without melody). Replicating the methods in Campfield and Murphy (2014), the emphasis during presentation was on listening and spoken production, with written forms presented as a visual, cross-modal support of oral language development (Jiang, 2025). Participants were asked to look at the screen, listen to the text and repeat it three times, and then each word or phrase was presented individually, with participants asked to repeat it. Using back-chaining to build up sentences from the end, words or phrases were added until the whole sentence was repeated, focusing on fluency and speed rather than meaning or forms. The content and the practice of repeating whole French sentences was new to the children, who were French beginners. The children quickly got used to the technique of listening and repeating, and were eager to join in the chorus of repetition. No puppets, hand gestures or actions were used during presentation, to avoid confounding the results with additional presentation methods (e.g., Chou, 2014, who confounded their results by presenting vocabulary through songs, storytelling with puppets, formal presentation by the teacher, and games).

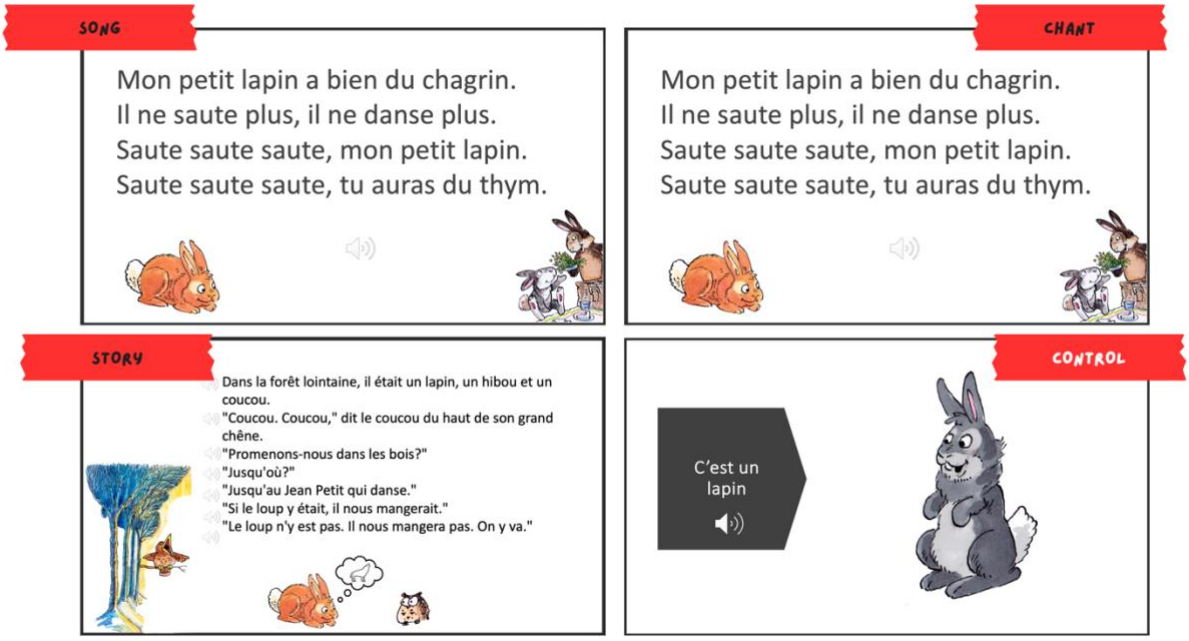
It was necessary to explain the outline of the story and premise of the song/chant materials in English, otherwise the children could not have engaged with the content. This explanation was not a line-by-line translation, and no focus on grammatical forms or word-by-word translation was provided. Key items of vocabulary such as the story's characters (*lapin*, *hibou*, *loup*) was given in English to assist the children in joining in with the narrative of the story and build rapport with the plight of the characters. For the song/chant materials, a quick explanation (e.g., "This is a song about the owner of a windmill who falls asleep and his mill starts going much too quickly!") sufficed to gain the children's interest and give them access to the overall meaning. Again, a few key items of vocabulary were translated (e.g., *trop vite* means *too fast* when explaining about the mill) but no word-for-word focused translations were provided. This was to give the children a more complex and implicit form of input, in contrast to the explicit and simplified content of the control condition, so that they had more exposure to the longer phrases entirely in French without interruptions for English translations.

Control

To match the experimental conditions in presentation style, the control materials were presented through a PowerPoint on large screen (as available in each teaching room). The same style of illustrations was used as in the experimental conditions, and the same typefaces and presentation of audio was pre-recorded and embedded in the presentation. Children were instructed to listen and repeat the audio in a chorus together, with the emphasis again on listening and speaking with additional visual support provided through the written forms. No writing was required. The teaching approach took a more explicit focus on grammar and translation in that I explained what each word meant in English, and how the sentence was constructed. For example, *C'est un chat* would be explained with "*C'est* means *this is* and *un* means *a* and *chat* means *cat*. So we are saying *This is a cat*. Let's try together: *C'est un chat*." As the animals changed but the phrase structure remained constant, the children built up their

ability to repeat *C'est un* more independently, and I would then focus on the pronunciation of the animals. When a feminine animal that takes *une* instead of *un* was encountered, I would explain that French has masculine and feminine words and give some examples. Hence there was more of a focus on small units of language, repeated structures with slot-filling variations, and grammatical forms, as appropriate for this age group.

Figure 4.15. Exemplar presentation materials for experimental and control groups



4.4.4 Primary outcome measure

This section outlines the rationale for choosing elicited imitation tasks (EIT) as the primary outcome measure. The first stage of parsing the research aim involved defining which aspect of beginner primary school French learners' linguistic outcomes would be considered for the primary outcome measure. Despite vocabulary knowledge being measured in 34 of the 60 included studies in Phase 1's systematic review, few reliable conclusions could be drawn from these studies about the influence of singing songs on L2 vocabulary uptake. This could be attributed to limitations in the methodology of the experiments, including designs that were not appropriate for claiming causal effects of singing on vocabulary acquisition, poor

accounting for confounders and biases in the designs, and statistical analyses that were ill-suited to model the collected data appropriately. Perhaps also there is a limitation in measuring the uptake of single lexical items presented through different conditions (e.g., song, chant, story). It could be the case that prosodically salient input (like songs and rhymes) has a more facilitative and detectable effect on learners' acquisition of phrase structure (Campfield & Murphy, 2013) due to providing more information about prosodic boundary cues and function words (Christophe et al., 1997) than single lexical items and thereby helping learners to parse the speech stream. Thus measuring emerging knowledge of phrase structure or multi-word items could prove to be more informative than pursuing evidence for knowledge of single lexical items, since the phrase structure may be acquired before the individual items (Wray, 2002). Indeed, the acquisition of English word order is investigated in Campfield and Murphy (2013), who found that prosodically salient input in the form of nursery rhymes assisted Polish 8–9-year-old EFL learners more than prose input in completing a grammaticality judgement task of English word order.

Paradoxically, however, this doctoral study was conducted with beginner French learners who had no prior formal knowledge of the French language, unlike Campfield and Murphy's (2013) participants who had been learning English for one year (54 hours) at the beginning of the studies. It would be challenging to measure my participants' French proficiency, since none could be assumed at baseline and there is a limit to how much they could learn in a short intervention period. It would also be demoralising to subject young learners to a test scenario that made them feel like they were failing at learning languages during the initial stages of their language learning journey. Careful consideration of the potential options for investigating the linguistic proficiency of these participants was therefore required to allow them to demonstrate progress, even within the short time frame of the intervention period, balanced with the level of challenge appropriate for novice learners. I decided that a grammaticality judgement task like the one in Campfield and Murphy (2013)

would be too challenging and produce data at floor that would not allow me to address the research aims adequately because participants in different groups would not be able to demonstrate their L2 competence through this measure.

If I was going to assess participants' acquisition of word order and build on the work of Campfield and Murphy (2013), I needed a measure that would clearly indicate when participants had produced the target word order correctly, even if they could not demonstrate explicit knowledge of it at such an early stage of their learning. I was restrained in choosing a target word-order feature by the consideration of which structures were available in the input, since participants could not be expected to acquire features they were not exposed to. The negative marker *ne […] pas* appeared six times in the input materials and is acquired early by French native speakers, emerging around the age of 2;1 years on average when children begin producing multiword units (Heinen & Kadow, 1990). It thus seemed a potential candidate for early acquisition by beginner French L2 learners. The placement of the negative marker differs in French and English, appearing after the verb (i.e., *ne [verb] pas*) in French, but between the main and auxiliary verbs in English (i.e., *[aux] not [verb]*). Bilingual French-English children rarely mix the word order placement for negative markers in either language (Paradis, Nicoladis & Genesee, 2000), suggesting that once acquired, this feature is represented in the morphosyntactic system pertaining to each grammar separately. Thus, if my participants had acquired the negative declarative word order in French, I hypothesised that they might correct test items containing a violation of this word order. Whilst a grammaticality judgement task would be too demanding of explicit knowledge at such an early stage in their learning, it might be possible to assess their implicit L2 competence regarding this feature since it appeared six times in the input.

There is some debate in the SLA literature about whether measures of L2 competence assess learners' implicit and contextualised L2 competence or their decontextualised, explicit knowledge of a surface-level L2 feature under controlled conditions (Norris & Ortega, 2000).

Doughty (2010) argues that even where instruction has been implicit, measures often assess explicit L2 knowledge in a discrete, decontextualised way through an explicit focus-on-forms approach. This decontextualised focus-on-forms assessment approach fails to tap into learners' L2 competence and performance, especially when learners are taught with implicit pedagogical methods, and this throws into question the construct validity of L2 instructional treatments and measures. Measures need to match the instruction, rather than measuring explicit knowledge when instruction has been implicit. One of the problems Doughty (2010) notes is that measuring learners' improvement in a particular aspect of the L2 often leads researchers to select an explicit measure, even if this does not match the instructional approach. This study aimed to overcome the issue of matching implicit instruction with explicit measures of L2 competence by using EIT as the primary assessment tool.

Elicited imitation is a test of participants' L1/L2 implicit language competence and can be applied to test their knowledge of specific grammatical structures, unlike free production tasks which cannot guarantee a target structure will appear in the speech sample (Erlam, 2006). Participants are asked to listen to an utterance (pre-recorded or read aloud) then to repeat it as exactly as possible. Gaillard and Tremblay (2016) argue that elicited imitation is a practical, reliable and valid tool for assessing L2 linguistic proficiency in the oral modality, since it reflects L2 processing efficiency. There is, however, some debate about whether EIT are reconstructive tasks or merely test rote imitation of stimuli (McDade, Simpson & Lamb, 1982; Vinther, 2002). If reconstructive, EIT participants process the stimuli and their elicited responses reflect the extent to which they decode and assimilate each stimulus into their internal grammar before reproducing an utterance (Munnich, Flynn & Martohardjono, 1994; Vinther, 2002). If merely testing participants' ability to repeat stimuli verbatim, then their working memory capacity will constrain their EIT performance on less familiar or more complex stimuli (Ellis, 2001; Gathercole & Baddeley, 1993) and only the sounds or acoustic image of the stimuli will be processed, without meaning being decoded (Vinther, 2002).

A first consideration for EIT design, then, is balancing the complexity, novelty and length of stimuli to create a test that measures participants' implicit L2 competence rather than their ability to repeat stimuli verbatim (Erlam, 2006). For young participants, it is expected that the ceiling level for performance on longer, more complex items will be quite low. In Campfield and Murphy (2014), a second study undertaken with 8–9-year-old Polish EFL learners, EIT practice items were 2–6 syllables long, and test items 4–9 syllables. Their results indicated that the rhythm-salient intervention group could imitate longer sentences more accurately and fluently, compared to the prose and control groups, and that 4–9 syllables presented enough of a challenge to prevent learners reaching ceiling performance. Presenting longer EIT sentences permits learners to demonstrate their full ability range, whereas making the test too easy could erroneously restrict nuances in the highest performances (Phakhiti, 2014).

Using an EIT in this study matched the assessment modes (speaking and listening) with the two primary input modes (speaking and listening), which received additional visual support from the written forms presented on the screen. It was hoped that this would reduce the problem of learners' L2 competence being misrepresented in assessment due to mismatching input and output modes (Murphy & Castillo, 2013) and also build on the prior work of Campfield and Murphy (2014) to see if their results could be replicated in a new context.

4.4.4.1 Elicited imitation task

The EIT used in this study replicates Campfield and Murphy (2014) with five practice and 22 test sentences. The practice sentences have one example each of sentences with 2–6 syllables. The test cue sentences have 4–9 syllables. No particular language structures were targeted. Rather, learners' developing general structural knowledge of French was assessed to ascertain whether exposure to rhythmically salient input, with or without the addition of a melody,

conferred any advantage compared to prose input or typical MFL classroom language input in the 'business as usual' control condition.

EIT stimuli (Table 4.5) were constructed to avoid any advantage of group membership. Every stimulus contained lexical and structural items that appeared across song, chant, and prose conditions, and from across the whole battery of input materials, thus not giving more advantage to input presented earlier or later in the three-week intervention period.

Table 4.5. EIT sources in experimental materials

Phase	Sentence ID	Sentence	Source (Song/chant & story instalment)
Practice	1	Pour nous.	Dansons la capucine & Story 5
	2	Que fais-tu?	Promenons-nous & Story 7
	3	Et puis encore	Sur le pont & Story 9
	4	Il nous mangerait.	Promenons-nous & Story 1
	5	Ton moulin va trop fort.	Meunier tu dors & Story 3
Test	1	Il sort sa tête	Petit escargot & Story 2
	2	Il ne saute pas	Mon petit lapin & Story 2/4
	3	Mon petit bouquin	Mon petit lapin & Story 7
	4	Ce n'est pour pas nous	Dansons la capucine & Story 10
	5	Remarquons les flots	Rame donc & Story 6
	6	Si le chat y était	Promenons-nous & Story 1
	7	Un kilomètre à pied	Un kilomètre à pied & Story 2
	8	Mais il n'y est pas comme	Promenons-nous & Story 11
	9	Du haut de son grand chêne	Dans la forêt lointaine & Story 1
	10	Il nous pas mangera	Promenons-nous & Story 1
	11	Savez-vous planter les roux?	Savez-vous planter les choux? & Story 5
	12	A la mode de chez nous	Savez-vous planter les choux? & Story 5
	13	On les plante avec les nez	Savez-vous planter les choux? & Story 5
	14	Y a chez nous pas de pain	Dansons la capucine & Story 8
	15	Le dindon n'a pas de nom	Petit poisson & Story 6
	16	Il n'avait jamais navigé	Un petit navire & Story 6
	17	Un éléphant se balançait	Un éléphant & Story 10
	18	Pendant que le loup n'y pas est	Promenons-nous & Story 7
	19	Trois petits tours et puis s'en vont	Ainsi font les marionettes & Story 10
	20	Petit chaton dis-moi ton nom	Petit poisson & Story 6
	21	Il était un petit navire	Un petit navire & Story 6
	22	Un petit poisson qui tourne en rond	Petit poisson & Story 6

Of the 22 test sentences, six were altered to include unfamiliar lexical items that maintain the same number of syllables and prosodic structure, and same rhymes as the familiar items (e.g. *bouquin* instead of *lapin*, *nez* instead of *pieds*). The aim was to test

whether learners could generalise their knowledge of L2 structure to unfamiliar contexts to provide data for answering RQ2c.

Additionally, three test sentences contain grammatical errors in the *ne[…] pas* clause (*Ce n'est pour pas* nous*; *Il nous pas* mangera*; *Pendant que le loup n'y pas* est*). French negation always has the verb inserted between *ne [verb] pas*, thus the word order in each case was varied at an invariable point to produce a grammatical error, not merely an infelicitous error. The inclusion of these grammatically incorrect items provided data for addressing RQ2d and permitted analysis of whether learners unconsciously corrected the grammatical errors in their utterance, and whether this varies across experimental groups. This provided insight into whether stimuli are processed, decoded and reproduced as utterances, or repeated verbatim. There were only three such items to avoid overburdening the young participants, but with 94 participants there were enough data points to permit inferential analyses (n = 282 per time point, 846 total). The changes from original input to stimuli on the EIT are shown in Table 4.6.

Table 4.6. EIT stimuli changes from input

Sentences marked with an asterix (*) include a novel item of vocabulary (sentences 2, 3, 5, 6, 11, 13, 15, 20) or a grammatical error (sentences 4, 10, 18), or a change in word order that is different from the input but grammatically correct (8, 14).

Phase	Sentence ID	Syllables	Sentence	Original input
Practice	1	2	Pour nous.	
	2	3	Que fais-tu?	
	3	4	Et puis encore	
	4	5	Il nous mangerait.	
	5	6	Ton moulin va trop fort.	
Test	1	4	Il sort sa tête	
	2	4	Il ne saute pas*	Il ne saute plus
	3	5	Mon petit bouquin*	Mon petit lapin
	4	5	Ce n'est pour pas* nous	Ce n'est pas pour nous
	5	5	Remarquons* les flots	Attaquons les flots
	6	6	Si le chat* y était	Si le loup y était
	7	6	Un kilomètre à pied	
	8	6	Mais il n'y est pas comme*	Mais comme il n'y est pas
	9	6	Du haut de son grand chêne	
	10	6	Il nous pas* mangera	Il nous mangera pas
	11	7	Savez-vous planter les roux*?	Savez-vous planter les choux?
	12	7	A la mode de chez nous	
	13	7	On les plante avec les nez*	On les plante avec les pieds
	14	7	Y a chez nous pas de pain*	Y a pas de pain chez nous
	15	7	Le dindon* n'a pas de nom	Le poisson n'a pas de nom
	16	8	Il n'avait jamais navigé	
	17	8	Un éléphant se balançait	
	18	8	Pendant que le loup n'y pas* est	Pendant que le loup n'y est pas
	19	8	Trois petits tours et puis s'en vont	
	20	8	Petit chaton* dis-moi ton nom	Petit poisson dis-moi ton nom
	21	9	Il était un petit navire	
	22	9	Un petit poisson qui tourne en rond	

4.4.4.2 Administration of EIT

The EIT was administered on a MacBook laptop, with sentences recorded as sound files and inserted into a PowerPoint presentation with the text appearing on screen (mirroring the presentation method during the intervention lessons, but without illustrations). The participant sat to the left of the researcher. They were presented with 'Blue' (a snowball microphone on a small desktop tripod with the word *Blue* on it) and a pair of headphones, both of which were connected to the laptop. The laptop faced the researcher. Participants were told that they were going to hear some French through the headphones and their job was to tell Blue what they could hear. They were then invited to put on the headphones and asked if they were ready to

begin with some short practice sentences. If they wanted to hear the sentence again, they were told they could just say 'Again' or 'Play it again' or gesture to that effect. One repetition per sentence was permitted. They were encouraged to have a go at saying any of the sounds they heard, with no expectation of getting it exactly right. Just to 'have a go' at saying what they heard. They were also told that they did not need to explain what the French sentences meant in English, just to repeat it as best as they could, even if that felt slightly odd as we do not usually repeat things we do not know the meaning of. The participants willingly and in many cases enthusiastically joined in with telling Blue what they could hear. One boy from School 1 did not wish to say anything, and after attempting to rationalise his fears and allow him time to get used to the procedure, he decided he wanted to stop, which we did.

Once a child had attempted to imitate a stimulus and indicated they were ready for the next sentence, we moved on. To initiate the next sentence, I pressed the enter button on the keyboard. The children understood that their responses (and the other children's) would be recorded to allow me to hear what they said later and try to match it up with what they heard through the headphones. I emphasised that they were not being tested on their knowledge of French words, as I was investigating how well they could perceive and repeat the sounds of words they did not know. They saw this as a fun game, although a few children did ask what the words meant and I explained that I would tell them all the meanings after we had finished the research project or it might compromise the experiment. As budding scientists, they were intrigued by the experimental process and seemed to enjoy being part of it.

The EIT was administered at pre- and immediate posttest, and after 6–7 weeks in a delayed posttest. The aim of the delayed posttest was to ascertain whether any differences between groups are maintained over a longer period, during which time participating class teachers were asked not to reuse any of the materials to avoid repetition effects. Given the strong intuition among practitioners that songs get 'stuck in our head' and confer a language-

learning advantage over other methods (Hamilton & Murphy, 2023), the delayed posttest was administered to provide some empirical evidence for inclusion in this discussion.

4.4.5 Testing the materials and procedure

Following ethical approval from CUREC, a pilot study was conducted in May–June 2023 to test all procedures including onboarding of schools, presentation of materials, and data collection and analysis methods. Having piloted the screening variables with 16 children, it was clear that the baseline tests (which also included the pretest) were taking too long with each child to be feasible once the number of participants increased to near 100. Therefore, the paper-based WASI test of verbal reasoning was replaced with the online PVST (Anthony & Nation, 2021) as described in 4.3.3.4, which meant that a whole group or class of children could be tested at once on this measure. Since English vocabulary knowledge was not the main focus of the experiment, the insight provided by the WASI vocabulary test into children's verbal reasoning seem unnecessarily detailed. All that was required is an indication of children's receptive vocabulary as a proxy measure of verbal ability. It was decided that the Matrix Reasoning test could be administered on paper as a quiet self-test exercise that children completed by circling in pencil their choice, as a class or group exercise rather than one-on-one. This would also help reduce the time taken to administer the screening variables to fit with the timeframe that schools were able to make available. The final refinement was to ask all children to wear the headphones whilst completing the EIT, rather than giving them a choice to listen through the laptop speakers, as then the sound quality was more stable for all participants. A full description of the pilot study is available in Appendix B6.

4.4.6 Administration of experiment

Following piloting in June 2023, the necessary adjustments and refinements were made to the materials and screening variables as described in 4.4.5. The main period of data collection

began in September 2023, continuing until the final delayed posttests in January 2024. Children were issued with a certificate of participation at the end of the study (Appendix B7).

4.5 Data analysis

4.5.1 Preparation for data analysis

Following data collection, data were prepared for analysis through several procedures to create one 'data collection' Excel spreadsheet for the screening variables, and one for the EIT measure. The EVIP scores were entered directly on the spreadsheet. The EVIP raw scores were used in group comparisons, since it is a test standardised on a French-speaking population and the standardised scores would not be applicable to this study's beginner French learner participants. The WASI Matrix Reasoning test standardised *t*-scores were calculated by looking up the *t*-score equivalent of each raw score in the age-level tables in the WASI test manual. These *t*-scores were then added to the data frame. The PVST was analysed in R using RStudio (Version 2023.12.1), the output of which produces a file with participant ID, School, number of correct items and vocabulary size statistic. The vocabulary size and number of correct items data were then added to the data frame for each participant. To calculate the c-RST scores, the provided Python code was run on a MacBook laptop. This code calculated the % correct taps per child, and their inter-tap interval ITI synchrony score, which were duly added to the data frame. Having gathered all four screening variables into one data frame, comparisons across the two schools and then the four experimental groups were carried out in RStudio, following the preregistered R script (Hamilton, Chalmers & Murphy, 2024).

For the EIT measure, data were first transcribed and then scored. In total there were 7668 sentences to process. To minimise bias when scoring, since it was the same person collecting the data and analysing the recordings, a third party (a proficient Excel user) generated an anonymised, randomised order for analysis. This involved each participant's audio-recordings from the three time points being assigned a random recording number, for example Rand175 could indicate a participant's pretest recording. A copy of all audio files

was made in a new folder in OneDrive, each renamed to the new random number to remove any links with time point or subject ID. Then a list of randomly ordered file name and sentence numbers (e.g., Rand175_Sen16) was created to follow for transcription and subsequent scoring. Scoring followed the ordinal 0–5 scale presented with examples in Table 4.7[16].

Transcriptions focused on capturing the sounds participants uttered in response to the stimuli. Utterances sometimes resembled the stimulus words enough to make transcription straightforward and sometimes presented challenges to capture accurately in phonemes. Where a sound did not have a direct correspondence with a phoneme from the stimulus, as far as possible French spelling conventions were used. Thus, for example, *ch* (as in *chien*) was used instead of the English *sh* [ʃ] in the transcription *Chuh fais-tu?* (in response to the stimulus: *Que fais-tu?*). In stimuli sentence 9 where both *son* and *chêne* were present, the sound corresponding to the *son* was rendered with *sh* if pronounced [ʃ] and the sound corresponding to *chêne* was rendered with the French *ch* to distinguish between them, and indicate that both sounds had been produced in response to different stimuli sounds (e.g., in the example *du tot de sho gro ché*). Transcriptions were written in Roman alphabet, rather than IPA, because IPA transcription was felt to be needlessly complex for the purposes of this study. The transcription and scoring process began after the final delayed posttests and pilot data analysis, on 29th January 2024, and (after a two-week intermission due to illness in February) was completed on 3rd April 2024. Between 150 and 300 sentences per day were transcribed and scored, in batches of 25 sentences. Some sentences took under one minute to transcribe, and some took several minutes, multiple listenings and slowing down of the playback. This was due in some cases to poor audio quality (a by-product of recording in a busy school environment, where noise is unavoidable even in a 'quiet' classroom), or to

[16] Following examiner feedback at the viva, I have added approximate IPA transcriptions to Table 4.7.

the children speaking very quietly or sometimes producing nearly unintelligible, unarticulated responses.

Given the subjectivity involved in transcribing and scoring the responses, a second rater who has a master's level qualification in applied linguistics and speaks fluent French was engaged to transcribe and score 10% (766) responses. I trained the second rater on a test batch of sentences, going over the scoring scale in detail and working through some examples together. The second rater was then given 766 randomly assigned and anonymised sentences so that neither rater knew the subject ID or time point of recording. When first reviewing the interrater reliability score, there was only 52% exact agreement between raters, which seemed rather low. There had been an incident with the second rater sorting the spreadsheet at the last minute and accidentally muddling the final 100 responses and ID numbers. However, this was rectified and carefully checked, and did not appear responsible for the low overall agreement. To check whether the issue lay with the sorting problem not having resolved itself, or whether there was an issue with training and use of the scale, the second rater transcribed and rated a second batch of 50 responses (25 previously rated and 25 unseen). Again, the exact agreement score was 52% precisely, indicating that the reviewers agreed on the precise score they gave only half of the time. It was noted that we agreed with plus or minus one mark 92% of the time in the batch of 766 responses, and 80% of the time in the batch of 50.

There could be several reasons for the low exact agreement and high near-agreement scores. Firstly, the audio quality was (as noted above) not always clear and, even when clear, the children themselves did not always speak clearly. Perceiving what the children uttered presented a subjective challenge. Sometimes, even when rating sentences together to check where discrepancies had arisen, a sound could have been perceived as a French feminine *e* ending (as in *saut<u>e</u>*) or as a schwa (as in *saut<u>a</u>*). This resulted in a rating of 5 (if perceived as correctly imitating *il ne saute pas*) or 4 (if perceived as producing *il ne sauta pas* with one phoneme error). Extensive discussion of this particular example did not resolve the issue: both

interpretations were arguably correct depending upon one's perception of the sound, which was borderline e/a. Given the time constraints and lack of funding for a doctoral project, it was not possible to engage a third rater to triangulate the scores. It was therefore decided, in discussion with supervisors, that a plus or minus one mark score of 92% interrater reliability was good enough to ascertain that both raters had applied the scale faithfully.

Having agreed the scoring for the EIT responses, a cumulative link mixed model was applied to model the data and produce the comparison of relative effects of the experimental conditions compared to the control group over time. The model development is explained in detail in section 4.5.2.

Table 4.7. EIT measure scoring scale

Score	**Description**	**Examples**					
0	Omission	No response	"too tricky"	"skip this one"			
1	*Single words, or nonsense only:* some correct/similar sounds, or snippets of correct prosody, but far off the mark; not enough syllables/too short or excessively long	di aut don gro hen /di o dɔ̃ gʁo hɛn/	du oh sa gr s /dy ǒ sa gʁ s/	il na ko so ché /il na ko so ʃe̯/			
2	*Poor imitation* but *showing evidence of phrase structure:* the prosody of the phrase has same/almost same number of syllables and intonation is approximately correct, even if the words initiated differ from the cue	di haut sé cha grah sié /di o se ʃa gʁɑ sje/	du yaut de gron son ché /dy jo də gʁon sɔ̃ ʃe̯/	il donc son grand ché /il dɔ̃k sɔ̃ gʁɑ̃ ʃe̯/	du haut a sco plon sté /dy o a sko plɔn ste/	du haut son gro ten /dy o sɔ̃ gʁo tɛn/	du haut sa gro chez no /dy o sa gʁo ʃe no/
3	*Understandable*, from which the cue sentence can be guessed. Correct prosody plus no more than 3 minor phoneme errors, or 2 words in wrong order, or the end-sounds of words missing as long as everything else is correct (prosody, word order, onset phoneme of words) so the sentence is correct except they chopped off the end of words	di haut de gron chon sen /di o də gʁɔ̃ ʃɔ̃ sɛn/	si haut de shon grand sene /si o də ʃɔ̃ gʁɑ̃ sɛnə/	du goh no so gro chene /dy go no so gʁo ʃɛn/	du tot de sho gro ché /dy to də ʃo gʁo ʃe̯/		
4	*Almost exact* imitation, no more than one word or phoneme error (or 1-phoneme insertion)	du haut de son grand chez /dy o də sɔ̃ gʁɑ̃ ʃe/	du haut de son gro chêne /dy o də sɔ̃ gʁo ʃɛn/	du l'haut de son grand chêne /dy lo də sɔ̃ gʁɑ̃ ʃɛn/			
5	*Exact* imitation	du haut de son grand chêne /dy o də sɔ̃ gʁɑ̃ ʃɛn/					

4.5.2 Data analysis plan

Having seen the limitations of the statistical models applied in many of the studies included in the systematic review, particularly the failure to address hierarchically structured repeated-measures data with appropriate models, I decided that a more robust model with mixed effects needed to be built to model the EIT response data. A cumulative link mixed model (CLMM) is an extension of logistic regression that can handle repeated measures and random effects when working with an ordinal outcome variable. It uses (typically) the logit link, a cumulative link function that models the cumulative probabilities of moving from one category (in the EIT from 0 > 1 > 2 > 3 > 4 > 5) to a higher category. Logit links have the formula:

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

Here p is the probability of an observation being in a specific category or higher, and $\left(\frac{p}{1-p}\right)$ is the odds of being in that category.

An alternative to logic links is probit links, an inverse normal link function. Probit links produce a coefficient that can be interpreted as the difference in Z-score associated with a one-unit (or one-category in this case) difference in the independent variable. Since probit is less intuitive to interpret, and produces only slight or non-existent differences in the overall model results to the logit link function, I decided to use the logit link.

The log-odds are useful in logistic regression because they transform the probability scale (which ranges from 0 to 1) to the log-odds scale (which ranges from negative infinity to positive infinity). This transformation allows for a sigmoidal (S-shaped) relationship between the predictors and the log-odds, making it suitable for modelling ordinal outcomes like the EIT scores. In the CLMM, the log-odds are used as the link function to model the relationship between predictor variables and the probability of an event (i.e., an increase in EIT score)

occurring. The model estimates coefficients for each predictor, indicating the impact on the log-odds of being in a higher EIT score category. Exponentiating the coefficients produces the odds ratios, representing how the odds of being in a higher category change for a one-unit change in the predictor. In this final respect, interpreting the output is not unlike interpreting interaction effects from a more familiar linear model with continuous variables.

Since the predictor variable in this study is categorical too, the final statistics show how a move from the reference (Control) group to the Song, Chant or Story group, affects the odds in each case of increasing the EIT score. If the odds ratio is greater than 1 and statistically significant (for this we check the *p*-value associated with the coefficient in the CLMM output), it suggests an increase in the odds of a better outcome for the Song, Chant or Story group respectively. Values < 1 indicate a negative association of moving from Control to experimental group, thus a decrease in the odds of scoring more highly on the EIT.

The two-part research questions motivate two similar but distinct statistical analyses. The first is a test of the interaction effects between the contrasts between the Control and the three experimental conditions (Song, Chant, Prose) and test times, and then the interaction effects between the contrasts between the three experimental conditions (Song and Chant, Song and Story) and test times. Therefore, two versions of the same model were fitted: once with contrasts between Control and the three experimental conditions respectively, and once with contrasts between Song and the other two experimental conditions (Chant and Story). I ran the minimal number of comparisons necessary to address the two parts of the research questions to minimise the chances of a type-1 error. Contrasts between Chant and Story were not fitted since Song is the focal condition of this research.

It is important to underline that the control condition is merely present as an experimental design control, not as a comparison teaching group. Since Control participants did not have access to the EIT stimuli in their input (as the experimental groups did), they could not be expected to learn them. It would not be a fair test to make this comparison and

infer any conclusions about the teaching methods. Control participants received the same amount of French exposure (in terms of minutes of French input), as the three experimental groups. If an experimental design control were not present, and no significant interaction effects were detected between the contrasts between the three experimental groups and between contrasts between test times, it would be impossible to disentangle null effects from a failure of the experiment to detect any effects at all.

The R script for the CLMM using the `ordinal` package in R (Christensen, 2023) was preregistered on OSF (Hamilton, Chalmers & Murphy, 2024), having been tested on the pilot data. Additionally, following feedback and discussion on the data analysis with colleague Professor Elizabeth Wonnacott, steps for calculating Bayes Factors were added to the script. This additional step (detailed in 4.5.2.1) permits conclusions to be drawn about whether there is evidence *for* the null hypothesis that there is an absence of an effect, something that cannot be determined from a non-significant *p*-value in a null hypothesis significance test, where absence of evidence *for an effect* does not indicate evidence of absence *of the effect*. Professor Wonnacott also suggested that, where appropriate, I might plot gains scores as a more interpretable set of outcome data because log-odds scores are notoriously difficult to understand, and my aim is to reach as many research consumers as possible with statistics that can be understood easily.

The final models are coded in R as follows. Firstly, the model which tests the interaction of Group*Time with contrasts between the Control and experimental conditions is coded:

```
fmm2 <- clmm(EIT_score ~ (Control_VERSUS_Song + Control_VERSUS_Chant +
Control_VERSUS_Story) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | ID) +
(1 + (Control_VERSUS_Song + Control_VERSUS_Chant + Control_VERSUS_Story) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

And the model testing the interaction of Group*Time with contrasts between the Song and Chant or Song and Story conditions is coded:

```
fmm3 <- clmm(EIT_score ~ (Song_VERSUS_Chant + Song_VERSUS_Story +
Song_VERSUS_Control) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | ID) +
(1 + (Song_VERSUS_Chant + Song_VERSUS_Story + Song_VERSUS_Control) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

The contrast between Song and Control appears again in this second model, but I have only interpreted the Song and Chant, or Song and Story contrasts and I did not contrast Chant and Story conditions to minimise the number of comparisons made.

The contrasts are coded using an R function (`lizcontrasts4`; Wonnacott & Viviani, n.d.) which employs a *simple coding scheme* (UCLA: Statistical Consulting Group, n.d.). The function creates two centred dummy variables which stand in place of a three-way factor and allow us to inspect each contrast separately against a reference level, much as we do for dummy-coded variables, as well as the interactions between these contrasts and other factors (in this case, test time). The centred coding means that other fixed effects in the model (e.g., syllables) can be evaluated as the average effects across all levels of the factor, and the intercept corresponds to the 'grand mean' (the mean of all cell means), whereas for dummy coding, the intercept corresponds to the cell mean of the reference group only.

Finally, again following examiner feedback from the Confirmation of Status viva, I reran the models taking into account the clustered nature of the data. Although they were randomly allocated on an individual basis to the four conditions, participants were taught in groups. There could therefore have been a 'group effect' where group dynamics may affect learning outcomes. To account for this nested data structure, I adjusted the models by adding in a Group/ID term to the first random effect as follows for the contrasts with Control as reference:

```
fmm2_nest <- clmm(EIT_score ~ (Control_VERSUS_Song + Control_VERSUS_Chant +
Control_VERSUS_Story) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | Group/ID) +
(1 + (Control_VERSUS_Song + Control_VERSUS_Chant + Control_VERSUS_Story) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

And then for the contrasts with Song as reference:

```
fmm3_nest <- clmm(EIT_score ~ (Song_VERSUS_Chant + Song_VERSUS_Story +
Song_VERSUS_Control) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | Group/ID) +
(1 + (Song_VERSUS_Chant + Song_VERSUS_Story + Song_VERSUS_Control) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

The results were almost identical to the original models without the nested Group/ID term. I therefore report the findings for the original models in Chapter 5, and include the output from the nested models in Appendix B8.2.

4.5.2.1 Bayes Factors

In the case where non-significant interaction effects are detected between experimental groups at posttest and delayed posttest based on the CLMM output, and since non-significant *p*-values cannot differentiate 'evidence for the null' from 'no evidence for any conclusion' (Fisher, 1935; Dienes, 2014), I additionally calculated Bayes factors (*B*). *B* can provide evidence of the extent to which observed data support the H_1 (that there is a difference between groups) versus the H_0 (that there is no difference between groups). To compute *B*, first a model of the H_1 is required which tells us the distribution of expected differences under the H_1. The H_1 being tested for research questions 2a, 2b and 2c is that there is an interaction between condition and test time, in particular that there is an interaction between the contrasts between conditions and between contrasts between test times. Since these are directional hypotheses (i.e., one-sided predictions), I used half-normal distributions in each *B* calculation. Because it is often harder to distinguish the alternative from the null with a half-normal distribution, when *B* does distinguish between the theories with a half-normal distribution then the conclusion is strengthened (Dienes, 2014) because it makes the calculation more conservative, and hence more robust.

Three numbers are required to calculate *B* following Dienes' (2008) calculator method[17]: the estimated mean difference in the data (1) and associated standard error (2), plus the estimate of the predicted mean difference under H_1 (3). For 1 and 2 I used the beta coefficient and associated standard error from the CLMM for each contrast (Song-Chant, or Song-Story) at posttest or delayed posttest. For 3, since no prior studies using the same methods exist, I used the Song-Control contrast beta because this provided a best estimate of maximum H_1 effect size for the observed data, divided by 2 (to give the half-normal).

[17] There are helpful expositions in Viviani, Ramscar and Wonnacott (2024) and appendices detailing the statistical analyses, plus R code to accompany Dienes' (2008) *B* calculator method in Baguley and Kaye (2010).

To interpret *B* we use a continuous scale from ≤ 1/100 to >100 that has discrete categories (Jeffreys, 1961; Dienes 2014) suggesting how strong the evidence is in support of H_1 or H_0. In the centre of the scale, *B* between 1/3 and ≤ 3 are inconclusive, supporting neither hypothesis unambiguously. Even though *B* can be interpreted continuously (unlike *p*-values), 1/3< *B* ≤ 3 should thus be interpreted cautiously. Categories with *B* <1/3 favour the H_0, and categories with *B* > 3 provide evidence for H_1. The more extreme the *B* value beyond <1/3 or >3, the more substantial the evidence for H_0 or H_1 respectively.

Since the predicted value used to inform H_1 is derived from the observed data and thus conclusions could depend upon this particular model of H_1, I calculated Robustness Regions (RRs) for each *B* (Silvey, Dienes & Wonnacott, 2021; Dienes, 2019). RRs provide the range [x:y] of distributions of H_1 under which *B* leads to the same conclusion about the model comparison (i.e., whether it indicates that H_1 or H_0 is consistently favoured across a range of prior distributions). In other words, x is the lowest and y the highest value that can be used to obtain *B* >3 (if *B* is >3) and lower than 1/3 (if *B* is lower than 1/3), or between 1/3 and 3 (if *B* is likewise 1/3< *B* ≤ 3).

Taken together, *B* and the RR indicate the strength and direction of the evidence provided by the data (in favour of the H_0 or H_1, or neither) and permit reliable conclusions to be drawn about whether there is evidence of absence of an effect (i.e., evidence for the H_0 in the absence of significant *p*-values from the CLMM), or simply absence of evidence either for H_1 or H_0.

4.5.3 Missing data and intention to treat

The chief scientific advantage of running a randomised controlled trial is minimising the bias associated with allocation of participants to conditions. The process of randomisation means that potential confounders (both known and unknown) are distributed on the basis of chance across groups. Any difference between the average characteristics of groups is a chance difference rather than a systematic difference (bias). All things being held equal, random

assignment to treatment conditions permits differences in the outcome variable between groups being assigned to an effect of the intervention, rather than to preexisting differences between groups. However, if participants do not complete the intended interventions according to protocol, the researcher must decide how to deal with the missing data (Bishop & Thompson, 2024).

In this study, two participants withdrew after randomisation had taken place, but before outcome data were collected, and were consequently excluded from data analysis (see 4.4.4). The final sample participants (n = 94) completed all outcome measures. However, as might be expected over a three-week period in the autumn term of school, there were some absences due to illness during the intervention period which meant that not all participants attended every intervention lesson. One option would be to run a 'per-protocol' analysis of the outcome data, including only those 56 participants who attended every intervention lesson. Per-protocol analysis, however, reintroduces bias at the analysis stage by disrupting the balance of confounding variables that randomly assigning participants to conditions had minimised. This can lead to "gross misinterpretation and inaccurate (biased) assessment of the effectiveness of the intervention" (McCoy, 2017: 1076). The problem with per-protocol analysis is that drop-outs tend to occur in a systematic way, and thus confound the randomisation of groups.

It is preferable, therefore, to analyse outcome data according to the groups to which participants were randomly assigned, even if they did not complete the intervention as intended. This is called an 'intention-to-treat' analysis (Bishop & Thompson, 2024) and it "preserves the prognostic balance afforded by randomisation" (McCoy, 2017: 1076). McCoy (2017) describes how participants who adhere to drug trials tend to have better outcomes regardless of whether they receive the intervention drug or the placebo. This is because there is an association between adhering to drug therapy and overall health behaviour that would lead to a positive outcome for patients. In the educational context, it would not entail a life-or-

death outcome as described in McCoy (2017), but there is a negative correlation between attainment and absenteeism (DfE, 2022; Klein, Sosu & Dare, 2022). By excluding absentees from the data analysis, bias is introduced into the group comparison that is unrelated to the effectiveness of the intervention. This could lead to misinterpreting group differences as an effect of the intervention, when differences may actually have occurred due to preexisting differences in non-randomised groups.

I conducted a sensitivity analysis by analysing the data from the 56 participants who had attended all intervention sessions. The outcomes showed a similar pattern for this subgroup as for the 94 participants in the whole sample (see Appendix B8). Findings are thus reported for the intention-to-treat analysis.

4.6 Ethics

4.6.1 The ethical warrant for the study

We do not currently have sufficient evidence that using songs to teach FL is an effective pedagogy for improving young learners' linguistic outcomes. As it stands, there is a lack of evidence arising from research studies where competing hypotheses have been robustly tested. Consequently teachers cannot rely upon research evidence to guide their pedagogical choices. Such decisions rely more on folk pedagogy (Bruner, 1996; Hamilton & Murphy, 2023) than robust evidence of the effects of using songs on substantive linguistic outcomes, as explored in Phase 1's systematic review. If practice decisions are made based solely on folk pedagogy (rather than a mixture of practical experience and robust research evidence), and such decisions have a potential effect on the linguistic outcomes of children learning languages, it could be the case that children do not learn as well as they might if presented with alternative FL pedagogy. The ethical response when facing uncertainty about competing hypotheses that might guide teachers' pedagogical choices is to test them empirically. Knowledge arising from such study will either confirm teachers' perceptions about songs' potentially valuable role in helping children learn FL, or bring alternative evidence to light to

help inform their choices. The uncertainty teachers face about whether songs are an effective pedagogy for teaching FL when the specific aim is to develop children's linguistic skills thus provides the ethical warrant for this investigation.

4.6.2 Informed consent

Once the experimental and outcome measure materials were created and the instruments for the screening variables chosen, ethical approval for the study was sought from the Central University Research Ethics Committee of the University of Oxford (CUREC). This section outlines the steps taken to meet the ethics requirements for conducting research in primary schools with YLLs. Since the study involved collecting audio data from children, opt-in consent from parents or guardians was required and obtained for all participants. This entailed parents reading an information sheet, and signing and returning the consent form to school in order for their child to take part. Full recruitment and consent procedures, including the information sheets and consent forms sent to parents, are outlined in Appendices B3–B4, following university research ethics guidelines for working with children. Verbal assent was gathered from children before every testing session, following the script in Appendix B9.
As participants were 7–8 years old, they may have found the assessment process tiring. To mitigate this, testing periods were kept to 30-minute sessions at pretest and 15 minutes at posttest and delayed posttest. All materials were age appropriate and required short verbal responses, pointing, or an online click-based response. Testing was designed to be interesting and not too arduous for the children.

Participants may have felt uncomfortable working one-to-one with a researcher. Therefore, the class teachers invited me in to meet the children and introduce the research project to the classes. We created a positive relationship between the teachers and myself, and the children came to see me as a visiting teacher as well as a researcher. I also took time to chat with and help settle the children when I collected them from class to do the baseline tests.

As a parent and experienced teacher, I endeavoured to put children at ease before beginning testing in the quiet rooms provided by the school. I introduced them to the laptop and microphone equipment, and explained how they would be used. I collected children's assent before each testing session and they could ask to discontinue testing if they wished once they had begun. I built a positive working relationship with the children during the three-week intervention period.

To avoid unnecessarily disrupting class activities, I liaised with the teachers to schedule testing at times that avoid children missing key lessons or repeatedly missing the same aspects of their school routine. This was achieved by me being flexible about when I was available throughout the day and having allowed contingency time in the schedule in case of unforeseen delays (such as a longer assembly, class trips, or forest school lessons that children did not want to miss). Some disruptions to the usual class activities were inevitable but by working closely with the teachers involved, it was possible to minimise the effects of taking children out of their usual classes to do another activity.

CUREC guidelines suggested video footage would have been more detail than I needed to collect to answer the research questions, and therefore unethical. I therefore had to rely on checking my own behaviours matched across conditions and schools, including remaining upbeat and lively in my delivery for all groups. I took notes after every day's sessions with observations about what had arisen during the teaching (such as children's participation levels or any disruptions to the lessons). Whilst my research journal cannot provide as reliable a document as a video to check my own delivery matched across the weeks and groups, it would not have been possible to video myself without capturing footage of the children too, thus this was not possible on ethical grounds.

4.6.3 Risk management and data protection

Participants' schools and parents were informed, through the relevant information letters and informed-consent process, that data would only be used for the purposes of this study and

ensuing publications, not for any other purpose; that data would not be shared with other organisations (within the limits of legality); and that data would be anonymised such that no individual could be identified from the data. Participants were assigned a unique identification number for the purposes of referring to them without using their names. Schools were referred to using a number as pseudonym. All data were stored on password-protected laptops and within the university's secure online repository that requires two-factor identification to gain access. Raw data were only shared within this repository and only with my supervisors. In reporting, no individuals could be directly identified. Participants could withdraw their consent to take part in the study without giving a reason, up to the point of data analysis. All of these measures to protect participants' data were explained before gathering informed consent and before any data collection began.

4.6.4 Reporting findings to participants

I committed to writing a summary of the study's results in plain language and making this available to participants who wished to have one. I also offered to present the results to interested staff at the two participating schools as part of their CPD twilight sessions.

The University of Oxford CUREC approved my ethics application (see approval email in Appendix B10).

Chapter Five

Phase 2: Results

5.1 Introduction

An analysis of research data gathered through Phase 2's elicited imitation task (EIT) is presented in this chapter, with the research questions posed in Chapter 4 reiterated and addressed. The results from each research question are presented in sequence. RQ2a is addressed by modelling the EIT data from pretest, posttest and delayed posttest using the cumulative link mixed model (CLMM) described in 4.5.2. The coefficients and log-odds output of the CLMM are reported. RQ2b and RQ2c are addressed using a subset of the EIT data for each question, again with relative effects of input condition modelled using the CLMM. For RQ2b, the subset of data comprised EIT stimuli that included language familiar from the input; for RQ2c the subset comprised stimuli with novel items of vocabulary. In addition to the CLMM, Bayes factors are also computed for RQ2a, RQ2b and RQ2c, testing the null hypothesis in each case using the beta coefficients and standard errors from the CLMM output for each question. RQ2d is addressed through an analysis of each group's rate of error correction in the EIT stimuli sentences containing the French negative grammatical structure.

5.2 Restatement of the research questions and methods of analysis

The intervention study addressed the following research questions:

RQ2: What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on:

a) all 22 stimuli?

b) a subset of fourteen previously encountered stimuli?

c) a subset of eight previously encountered stimuli containing novel vocabulary items?

d) their ability to correct a grammatical error in the *ne [verb] pas* negative word order in a subset of three previously encountered stimuli?

To address RQ2a, b and c, the EIT outcome data were modelled with a CLMM which calculated the cumulative probabilities of moving from one category (in the EIT from 0 > 1 > 2 > 3 > 4 > 5) to a higher category. Two versions of the model were fitted, corresponding to two parts of the research questions with effects of the input presented in the three experimental conditions compared to i) the control and ii) each other. Having ascertained that the model fitted the data by running the CLMM with Control contrasted with the three experimental groups, I ran the second version of the model with Song as the reference group to get the relative effects of the experimental conditions (see section 4.5.2 for details). The effect of interest is the interaction between the contrasts between conditions and between contrasts between test times.

The two versions of the CLMM (with the two versions of contrasts) were used to address RQ2b and RQ2c, each with a subset of the EIT outcome data. For RQ2b, only responses to stimuli containing familiar items (i.e., with no change from the input) were included. This subset included fourteen responses per participant (items 1, 4, 7, 8, 9, 10, 12, 14, 16, 17, 18, 19, 21, 22), totalling 3934 observations. For RQ2c, the subset included only items with novel vocabulary (i.e., items that maintained the same prosody as the input vocabulary items but were not encountered in the input, see section 4.4.4.1). This subset included eight responses per participant (items 2, 3, 5, 6, 11, 13, 15, 20), totalling 2248 observations.

To investigate whether any non-significant Group*Time interactions for these three research questions were small enough to provide evidence *for* the null hypothesis that there is no effect in each instance, I also calculated Bayes factors for RQ2a, RQ2b and RQ2c.

To address RQ2d, responses to three stimuli that included a grammatical error (items 4, 10 and 18) in the *ne [verb] pas* construction were coded '0' if the error was repeated and '1' if the error was corrected. A correction entailed moving the *pas* to the correct place in the word order, so that the sentence followed the correct *ne [verb] pas* construction for French negation. There were 851 responses in the subset of sentences containing grammatical errors. The total number of accurately corrected responses is reported.

All analyses were conducted using R version 4.3.3 (2024) in RStudio version 2023.12.1+402 (RStudio Team, 2023) on an iMac late 2015 running OS Monterey version 12.7.4.

5.3 Data analysis

5.3.1 Research Question 2a

What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on all 22 stimuli?

5.3.1.1 Descriptive statistics of the EIT results

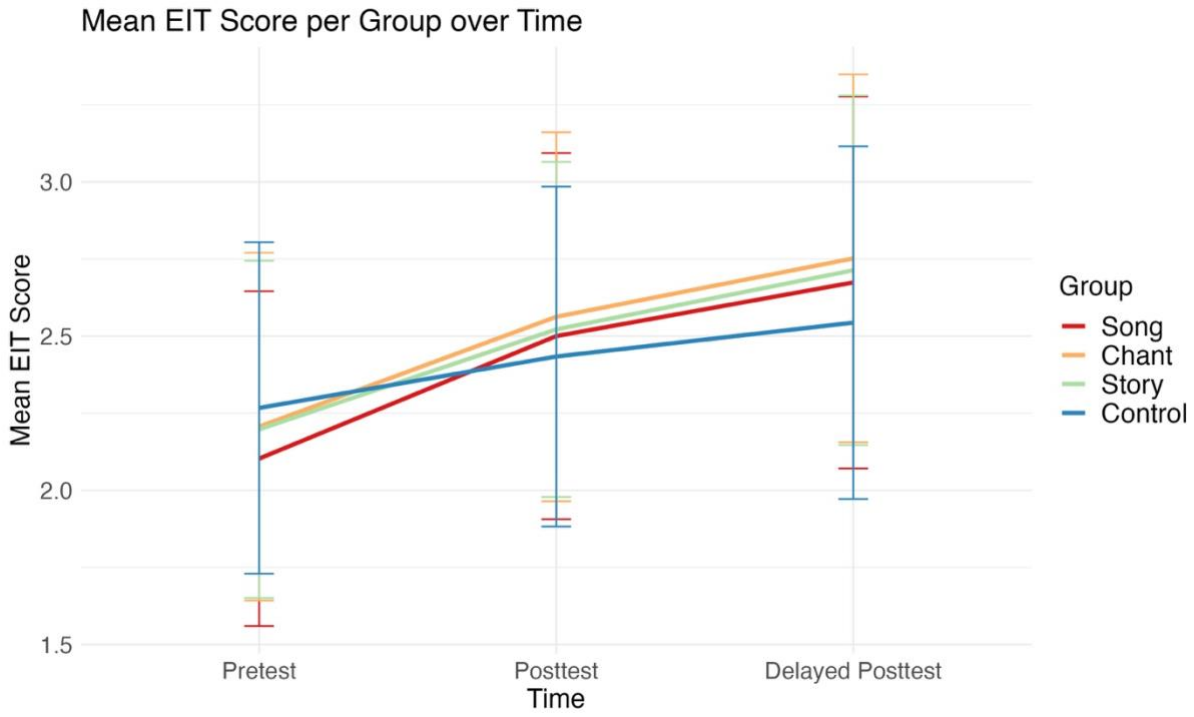
There were 7668 total responses from 94 participants on the EIT outcome measure across the three time points. Table 5.1 presents the mean and standard deviations of group EIT outcome

data at pretest, posttest and delayed posttest. Visual inspection of mean scores in Figure 5.1 revealed what appeared to be a Group*Time interaction effect. Please note the truncated Y-axis, which exaggerates the differences between groups but makes it easy to see the pattern of the trajectories and the potential interaction effect.

Table 5.1. Descriptive statistics for RQ2a

Group	Time	Mean	SD	SE	Lower 95% CI	Upper 95% CI
Song	Pretest	2.10	1.33	0.28	1.56	2.65
	Posttest	2.50	1.45	0.30	1.91	3.09
	Delayed posttest	2.67	1.48	0.31	2.07	3.28
Chant	Pretest	2.21	1.40	0.29	1.64	2.77
	Posttest	2.56	1.46	0.31	1.96	3.16
	Delayed posttest	2.75	1.50	0.30	2.16	3.35
Story	Pretest	2.20	1.34	0.28	1.65	2.74
	Posttest	2.52	1.33	0.28	1.98	3.06
	Delayed posttest	2.71	1.39	0.29	2.15	3.28
Control	Pretest	2.27	1.32	0.27	1.73	2.80
	Posttest	2.43	1.35	0.28	1.88	2.98
	Delayed posttest	2.54	1.40	0.29	1.97	3.11

Figure 5.1. Line plot of mean EIT score per group over time with error bars showing 95% CI (RQ2a)



5.3.1.2 Model fit and diagnostics

To check model fit, I first ran a cumulative link model (CLM) with fixed factors coded in R as follows:

```
clmm(EIT_score ~ (Control_VERSUS_Song + Control_VERSUS_Chant +
Control_VERSUS_Story) * (T1_VERSUS_T2) + syllablesen
```

I also ran a CLMM with fixed *and* random factors, coded in R as follows:

```
clmm(EIT_score ~ (Control_VERSUS_Song + Control_VERSUS_Chant +
Control_VERSUS_Story) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
               (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | ID) +
(1 + (Control_VERSUS_Song + Control_VERSUS_Chant + Control_VERSUS_Story) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

Then, to test the assumption of proportional odds for cumulative link models, I conducted a likelihood ratio test comparing the CLM and CLMM. Table 5.2 shows the model comparison statistics. The highly significant likelihood ratio result (*LR.stat* = 2410.2, *df* = 84, $p < .001$), indicated that the CLMM provided a much better fit to the data than the CLM.

Table 5.2. Model comparison (RQ2a)

Model	**No. of Parameters**	**AIC**	**Log-Likelihood**
CLM	17	17664	-8815.1
CLMM	101	15422	-7610.0

5.3.1.3 Results of the CLMM for RQ2a(i)

The CLMM was used to test the interaction between the contrasts of the three experimental conditions and control group with the contrasts between test times. There was a main effect of time, with all groups improving their scores at posttest (*b* = 0.65, *SE* = 0.11, *z* = 5.97,

$p < .001$) and delayed posttest ($b = 0.999$, $SE = 0.11$, $z = 9.44$, $p < .001$). There was a main effect of syllables, with scores decreasing across all groups as the EIT stimuli increased in length ($b = -1.41$, $SE = 0.14$, $z = -10.27$, $p < .001$).

RQ2a(i): Experimental conditions compared to control

All experimental conditions scored descriptively more highly than Control at posttest and delayed posttest. There was a significant interaction between the contrast between Song and Control and between pretest and posttest, reflecting greater change in the song condition than in the control group ($b = 0.52$, $SE = 0.23$, $z = -2.25$, $p = .024$). The odds of score increase for Song were 1.68 times those of Control from pretest to posttest, suggesting that participants in the song condition showed greater improvement in performance over this period. The interactions between the contrasts between Control and Chant ($b = 0.36$, $SE = 0.23$, $z = 1.60$, $p = .11$) and Control and Story ($b = 0.42$, $SE = 0.25$, $z = 1.69$, $p = .092$) and between pretest and posttest did not reach statistical significance.

The interaction between the contrasts between Song and Control and between pretest and delayed posttest was statistically significant ($b = 0.72$, $SE = 0.24$, $z = 2.95$, $p = .0032$), indicating that the odds of improvement for Song participants from pretest to delayed posttest were 2.05 times those of Control. The interaction between the contrasts between pretest and delayed posttest and between contrasts between Control and Chant ($b = 0.57$, $SE = 0.23$, $z = 2.51$, $p = .012$) and Control and Story ($b = 0.59$, $SE = 0.24$, $z = 2.43$, $p = .0152$) reached statistical significance, with higher odds of improving their EIT scores (Chant $OR = 1.77$; Story $OR = 1.81$) of approximately the same magnitude as Song ($OR = 2.05$; i.e., about twice as likely to receive a higher score on the EIT as Control participants).

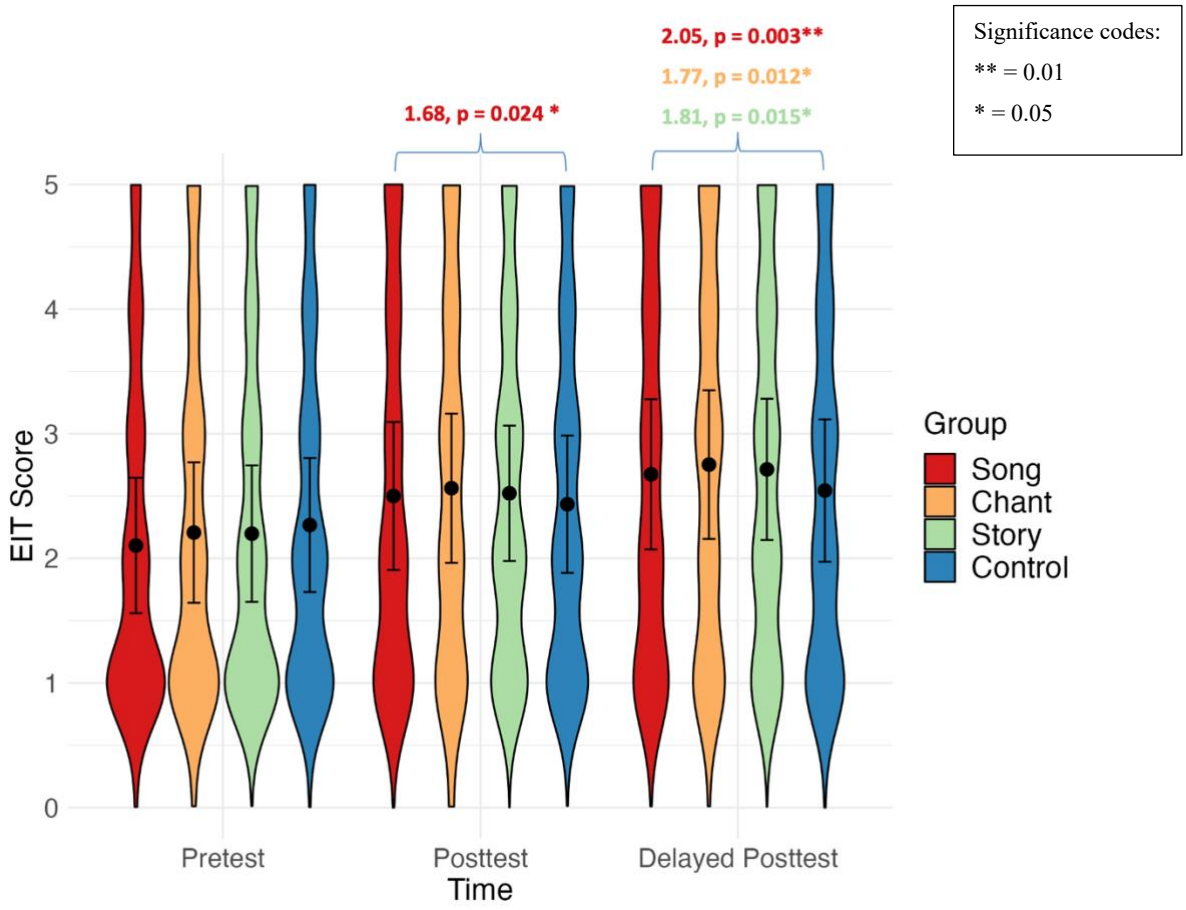
RQ2a(ii): Relative effects between experimental conditions

Neither of the interactions between contrasts between Song and Chant and between pretest and posttest (b = -0.15, SE = 0.24, z = -0.64, p = .52, OR = 0.86), or between pretest and delayed posttest (b = -0.15, SE = 0.24, z = -0.62, p = .534, OR = 0.86) were statistically significant. The picture is similar for the comparison of Song to Story. There was no statistically significant interaction between contrasts between Song and Story and between pretest and posttest (b = -0.097, SE = 0.25, z = -0.39, p = .694, OR = 0.91), or between pretest and delayed posttest (b = -0.13, SE = 0.23, z = -0.55, p = .585, OR = 0.88).

To minimise the risk of Type 1 error associated with conducting multiple statistical tests, I did not test the interaction between the contrast between Chant and Story and between pretest and posttest, or between pretest and delayed posttest. The interaction effects could be assumed to be non-significant as they were smaller than between contrasts between Song and Story and pretest and posttest, or pretest and delayed posttest. Relative to each other, then, the gathered data does not indicate an interaction effect between contrasts between experimental conditions and between test times.

Figure 5.2's violin plot visualises the distribution of EIT scores (0 to 5) in each condition at each of the three test times, with the mean score and confidence intervals indicated with the dots and whiskers, and with the statistically significant odds ratios superimposed. The violin plot provides a clear indication of how close the distribution of scores is between groups, in comparison to the line plot in Figure 5.1 where the Y-axis truncation exaggerates the differences between groups.

Figure 5.2. Violin plot showing EIT scores, significant ORs, group means and 95% confidence intervals (RQ2a)



In summary, CLMM analyses did not detect any statistically significant interactions between the contrasts between experimental conditions and between contrasts between pretest and posttest, or pretest and delayed posttest. Only the interactions between contrasts between Control and Song and between pretest and posttest, and between contrasts between Control and all three experimental conditions respectively and between pretest and delayed posttest reached statistical significance. The H_0 stated "there is no interaction between input condition and time on participants' overall performance on the elicited imitation task." Based on the results of the CLMM analyses, we can reject only the H_0 that there is no interaction between input condition and time in the case of Control and Song at posttest and Control and all three experimental conditions at delayed posttest. We cannot reject the H_0 that there is no

interaction between experimental conditions and time. The CLMM analyses do not provide evidence *for the null,* thus we cannot state there is no interaction between experimental conditions and time based on these analyses, only that there is no evidence of statistically significant interactions between experimental conditions and time on the EIT outcome measure.

5.3.1.4 Bayes Factors for RQ2a

Since the CLMM produced only non-significant *p*-values and thus no evidence either for H_1 or H_0, I calculated Bayes factors (*B*; see 4.5.2.1 for definition and rationale). *B* computation (following the calculator method from Dienes, 2008) requires three values: the estimated mean difference in the data (1) and associated standard error (2), and the estimate of the predicted mean difference under H_1 (3). The CLMM outputs provide all of these values as outlined in the following description of the calculations.

To test the Group*Time interactions of the song and chant, and song and story conditions at posttest and delayed posttest, I computed *B* to test one-tailed predictions of Song against Chant and Song against Story at each time point. The H_1 for each contrast tested at each time point were as follows:

1) Participants in the song group are predicted to score more highly than participants in the chant group at a) posttest and b) delayed posttest.
2) Participants in the song group are predicted to score more highly than participants in the story group at a) posttest and b) delayed posttest.

For the estimate of the predicted mean difference under H_1 (3), I used the contrast between Control and Song at each time point to provide the maximum difference that is possible from the observed data, because that was the contrast with the greatest magnitude. This value was

b = 0.51909 at the posttest and b = 0.718373445 at delayed posttest. I expected an effect in this region but that smaller values were more likely than large ones. Following Dienes (2008; 2021), I modelled H_1 as a half normal (mode = 0 and SD set to the predicted beta values for posttest and delayed posttest above). These values were halved for each B calculation because of the directional hypotheses that performance for Song is likely to be greater than Chant or Story at posttest and delayed posttest respectively. The prediction is set to be half of the size of the maximum, on the basis that the maximum of the normal distribution is about two SD (Dienes, 2008; 2021). Thus, we get the maximum plausible difference under H_1 to test the contrasts between Song and other experimental groups each time point.

For **hypothesis 1a**, there was no evidence either way of a difference between Song and Chant at posttest, M_{diff} = .06, b = -0.15, SE = 0.24, p = .52, $B_{HN(0, .26)}$ = 0.48, $RR_{B<1/3}[0.5{:}\infty]$.

For **hypothesis 1b**, there was no evidence either way of a difference between Song and Chant at delayed posttest, M_{diff} = .08, b = -0.15, SE = 0.24, p = .534, $B_{HN(0, .36)}$ = 0.38, $RR_{B<1/3}[0.5{:}\infty]$.

For **hypothesis 2a**, there was no evidence either way of a difference between Song and Story at posttest, M_{diff} = .02, b = -0.098, SE = 0.25, p = .694, $B_{HN(0, .26)}$ = 0.56, $RR_{B<1/3}[0.6{:}\infty]$.

For **hypothesis 2b**, there was no evidence either way of a difference between Song and Story at delayed posttest, M_{diff} = .04, b = -0.13, SE = 0.23, p = .585, $B_{HN(0, .36)}$ = 0.39, $RR_{B<1/3}[0.5{:}\infty]$.

To summarise, B calculations to test the one-tailed hypotheses that Song is predicted to score more highly than Chant or Story at posttest or delayed posttest returned ambiguous evidence that slightly favoured the H_0. However, since all B and RR reported above for RQ2a indicate that the data is insensitive ($1/3 < B \leq 3$), hence unable to distinguish between H_1 and

H_0, the conclusion is that this data does not provide evidence of no difference between Song and Chant or Song and Story at the two posttests. Overall, for RQ2a, the data indicate no significant Group*Time interactions according to the CLMM output. When testing the H_0 for evidence of no effect, however, Bayes factors were unable to provide anything more than ambiguous support: more data would be required to ascertain whether there is substantial support for the H_0 (B <1/3).

5.3.2 Research Question 2b

What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on a subset of fourteen previously encountered stimuli?

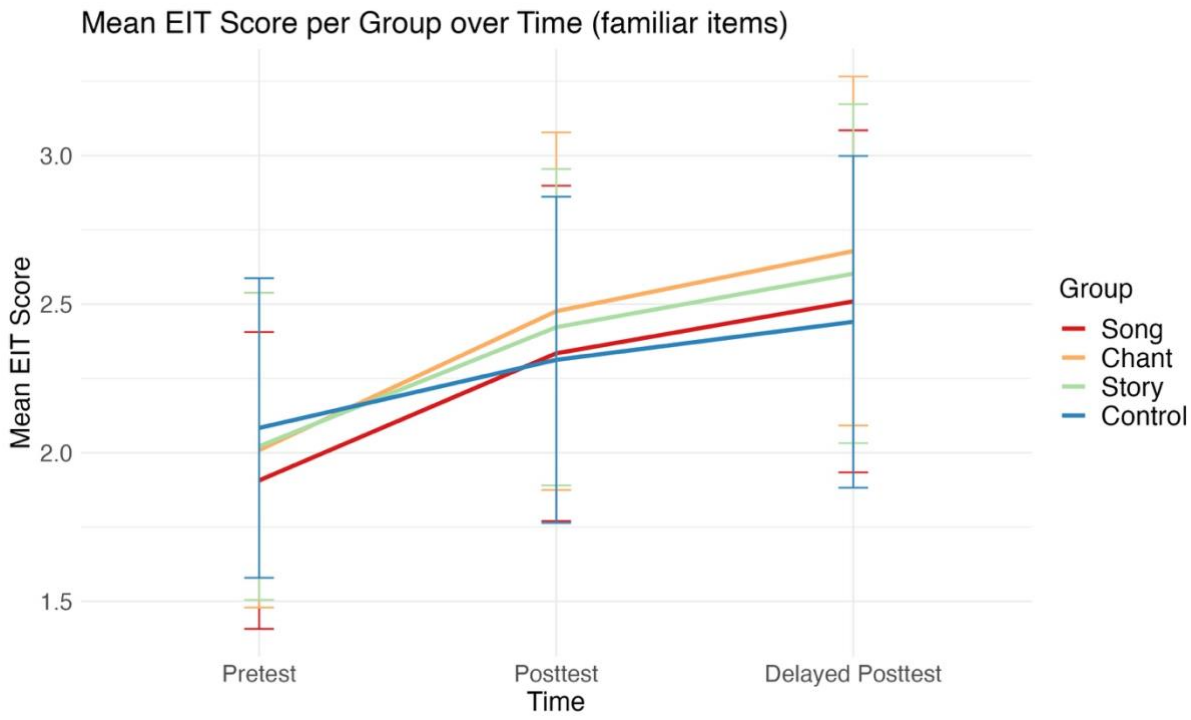
5.3.2.1 Descriptive statistics for RQ2b

There were 3934 responses to stimuli that were familiar from the input. The mean and standard deviation for each group at each time point are summarised in Table 5.3. Figure 5.3 visualises the mean score of each group over time in a line plot. Please note again the truncated Y-axis, which exaggerates the mean differences between groups but makes it easier to see the potential interaction of Group*Time.

Table 5.3. Descriptive statistics for RQ2b

Group	Time	Mean	SD	SE	Lower 95% CI	Upper 95% CI
Song	Pretest	1.91	1.22	0.25	1.41	2.41
	Posttest	2.33	1.38	0.29	1.77	2.90
	Delayed posttest	2.51	1.41	0.29	1.93	3.08
Chant	Pretest	2.01	1.30	0.27	1.48	2.54
	Posttest	2.48	1.47	0.31	1.87	3.08
	Delayed posttest	2.68	1.44	0.30	2.09	3.27
Story	Pretest	2.02	1.26	0.26	1.51	2.54
	Posttest	2.42	1.30	0.27	1.89	2.95
	Delayed posttest	2.60	1.40	0.29	2.03	3.17
Control	Pretest	2.08	1.23	0.26	1.58	2.59
	Posttest	2.31	1.34	0.28	1.76	2.86
	Delayed posttest	2.44	1.37	0.28	1.88	2.99

Figure 5.3. Line plot of mean EIT score on familiar items per group over time with 95% CI (RQ2b)



5.3.2.2 Model fit and diagnostics

With the smaller dataset for RQ2b, the CLMM converged with Control as reference group, but not with Song as reference. This suggested that the model was potentially too complex for

the reduced dataset, or there was multicollinearity between the predictors when Song was the reference group.

I checked for multicollinearity between the predictor variables by calculating the variance inflation factor (VIF) for each predictor using the `vif()` function from the `car` package in R. The linear model used was as follows:

```
lm(EIT_score ~ Song_VERSUS_Chant + Song_VERSUS_Story +
Song_VERSUS_Control, data = df)
```

VIF values under five suggest that there is no problematic multicollinearity between predictors (Kutner, 2005). The VIF values for the predictors were Song–Chant (1.53), Song–Story (1.52), and Song–Control (1.53). These values indicated that there was some correlation between the predictors, but not enough to cause the CLMM's failure to converge. I next confirmed that the data did not contain any blank values or errors. I then concluded that the model was too complex to use with the reduced dataset and rebuilt it, beginning with the fixed factors, and then adding the first random effect `(1 + Time|ID)`, and then `(1|Sentence.ID)`. This simpler model (without the random effects interaction) converged successfully and produced values that followed a similar trend to both the descriptive statistics and the CLMM with Control as reference group. The model with two random effects fitted the data much better than the CLM with only fixed factors, as indicated by the highly significant likelihood ratio result (*LR.stat* = 1406.7, *df* = 7, $p < .001$).
Table 5.4 shows the model comparison statistics for the CLM and CLMM with Song as reference group.

Table 5.4. Model comparison RQ2b

Model	**No. of Parameters**	**AIC**	**Log-Likelihood**
CLM	17	11166.5	-5566.3
CLMM	24	9773.9	-4862.9

5.3.2.3 Results of the CLMM for RQ2b(i)

I used a CLMM to test the interactions between the contrasts between test times and between conditions. There was a main effect of time, with all groups improving their scores at posttest (b = 0.81, SE = 0.14, z = 5.94, p < .001) and delayed posttest (b = 1.17, SE = 0.13, z = 8.97, p < .001). There was a main effect of syllables, with scores decreasing in all groups as the EIT stimuli increased in length (b = -1.21, SE = 0.18, z = -6.88, p < .001).

All experimental conditions scored descriptively more highly than Control at posttest and delayed posttest. There were no statistically significant interactions between the contrasts between pretest and posttest and between the contrasts between Control and Song (b = 0.47, SE = 0.29, z = 1.58, p = .114, OR = 1.59); Control and Chant (b = 0.52, SE = 0.28, z = 1.85, p = .064, OR = 1.68); or Control and Story (b = 0.41, SE = 0.32, z = 1.28, p = .202, OR = 1.51). There were significant interactions between the contrasts between pretest and delayed posttest and the contrasts between Control and Song (b = 0.65, SE = 0.32, z = 2.055, p = .0399, OR = 1.91) and Control and Chant (b = 0.68, SE = 0.27, z = 2.51, p = .012, OR = 1.97), but not Control and Story (b = 0.55, SE = 0.32, z = 1.74, p = .081, OR = 1.73).

RQ2b(ii): Relative effects between experimental conditions

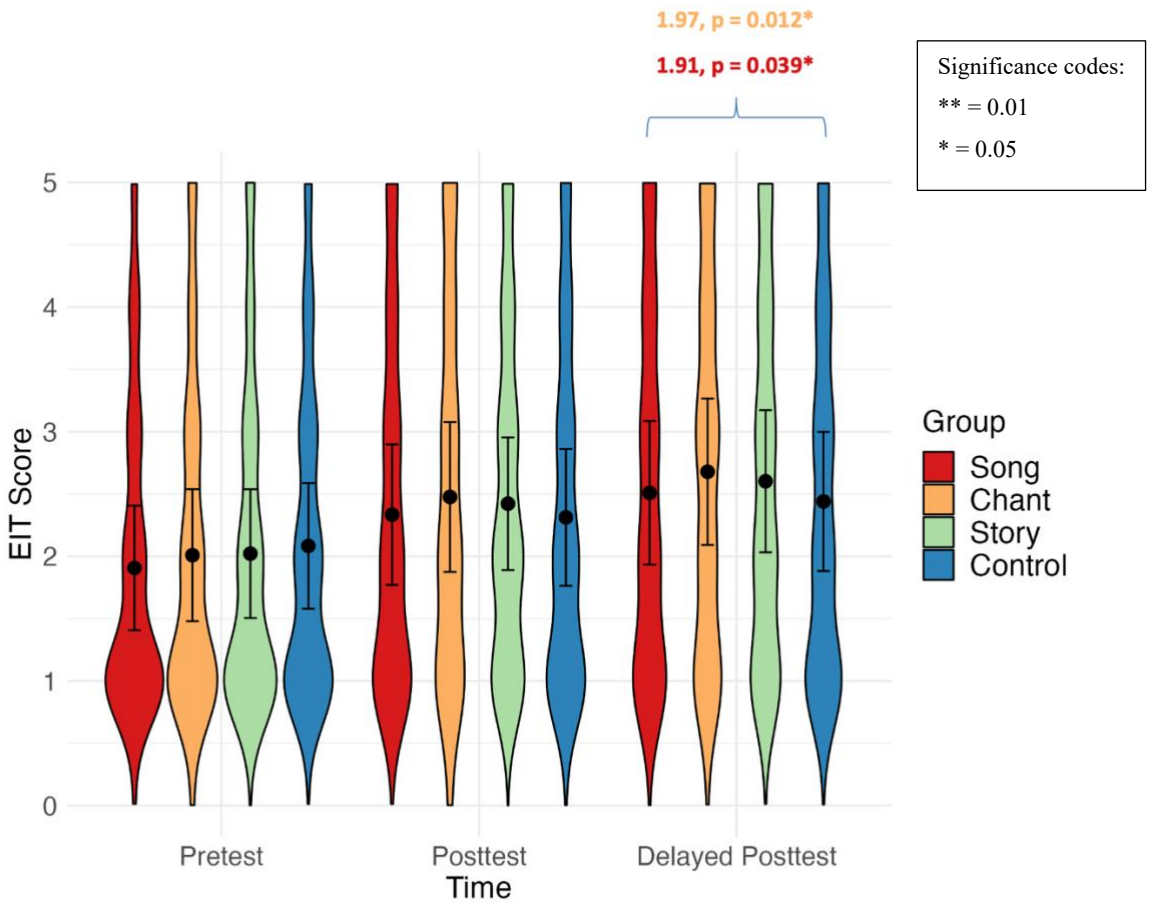
Regarding the comparison between experimental conditions on items familiar from the input, there were no statistically significant interactions between contrasts between experimental conditions and contrasts between test times as shown in Table 5.5.

Table 5.5. Relative effects between experimental groups on familiar items RQ2b(ii)

Group/time	**Estimate**	**SE**	***z*-score**	***p*-value**	**Odds ratio**
Song and Chant * Posttest	0.07	0.24	0.28	0.7778	1.07
Song and Chant * Delayed posttest	0.03	0.25	0.12	0.904	1.03
Song and Story * Posttest	-0.07	0.24	-0.28	0.78	0.93
Song and Story * Delayed posttest	-0.12	0.25	-0.47	0.631	0.89

The relative effects of the experimental conditions on participants' ability to reproduce familiar items indicate very small differences between Song, Chant and Story at posttest and delayed posttest. Figure 5.4's violin plot visualises the distribution of EIT scores (0 to 5) on familiar items in each condition at each of the three test times, with the mean score and confidence intervals indicated with the dots and whiskers, and with the statistically significant odds ratios between Song and Control, and Chant and Control at delayed posttest superimposed.

Figure 5.4. Violin plot showing EIT scores on familiar items, significant ORs, group means and 95% confidence intervals (RQ2b)



In summary for RQ2b(ii), there were no statistically significant interactions between the contrasts between pretest and posttest or delayed posttest and contrasts between Song and Story or Song and Chant. Only the interactions between the contrasts between pretest and delayed posttest and between the contrasts between Control and Song, or Control and Chant were statistically significant. The H_0 stated "there is no interaction between input condition and time on participants' performance on previously encountered stimuli in the elicited imitation task." Based these results, there is insufficient statistical evidence to reject the H_0 except in the case of the interaction between Song and Chant conditions compared to Control at the delayed posttest. Experimental input condition does not appear to predict participants' ability to reproduce previously encountered language in the elicited imitation task. Whilst the

mean differences between experimental groups are very small, the CLMM analyses presented here do not provide evidence *for the null.* We thus cannot state there is no difference between experimental conditions based on the CLMM output, only that there is no evidence of statistically significant interactions between test time and experimental conditions relative to each other, or between test time and Story and Control.

5.3.2.4 Bayes Factors for RQ2b

Given that *p*-values cannot distinguish between absence of evidence, and no evidence of absence when testing hypotheses, I calculated Bayes factors (*B*) for each of the following hypotheses for RQ2b:

1) Participants in the song group are predicted to score more highly than participants in the chant group at a) posttest and b) delayed posttest.
2) Participants in the song group are predicted to score more highly than participants in the story group at a) posttest and b) delayed posttest.

I ran similar *B* analyses to RQ2a, testing one-sided predictions and therefore using a half-normal distribution for H_1 following Dienes (2008, 2021). For the estimate of the H_1 predicted mean difference, I used the contrast between Chant and Control, since this was the largest effect in the CLMM for RQ2b. The beta value was 0.51847169 at posttest and 0.67694455 at delayed posttest.

For **hypothesis 1a**, there was no evidence either way of a difference between Song and Chant at posttest, $M_{diff} = .15$, $b = 0.07$, $SE = 0.24$, $p = .78$, $B_{HN(0, .26)} = 0.81$, $RR_{B<1/3}[0.9{:}\infty]$.

For **hypothesis 1b**, there was no evidence either way of a difference between Song and Chant at delayed posttest, M_{diff} = .17, b = 0.03, SE = 0.25, p = .904, $B_{HN(0, .34)}$ = 0.63, $RR_{B<1/3}[0.8:\infty]$.

For **hypothesis 2a**, there was no evidence either way of a difference between Song and Story at posttest, M_{diff} = .09, b = -0.07, SE = 0.24, p = .78, $B_{HN(0, .26)}$ = 0.58, $RR_{B<1/3}[0.6:\infty]$.

For **hypothesis 2b**, there was no evidence either way of a difference between Song and Story at delayed posttest, M_{diff} = .09, b = -0.12, SE = 0.25, p = .631, $B_{HN(0, .34)}$ = 0.44, $RR_{B<1/3}[0.5:\infty]$.

In summary, Bayes factor calculations testing the one-tailed prediction that the song group is more likely to score highly than the chant or story groups on familiar items in the EIT at posttest or delayed posttest found that the data is insensitive. All four B were between 1/3 and ≤3, indicating no evidence either for H_1 or H_0. For RQ2b overall, then, the CLMM found no significant Group*Time interactions, but testing the interactions between song and chant or story groups with Bayesian analyses also found no substantial evidence of absence of effect. It is not possible, therefore, to state that the experimental groups are equally likely to improve their performance on the EIT measure with familiar items, just that they did not differ statistically significantly from one another.

5.3.3 Research Question 2c

What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on a subset of eight previously encountered stimuli containing novel vocabulary items?

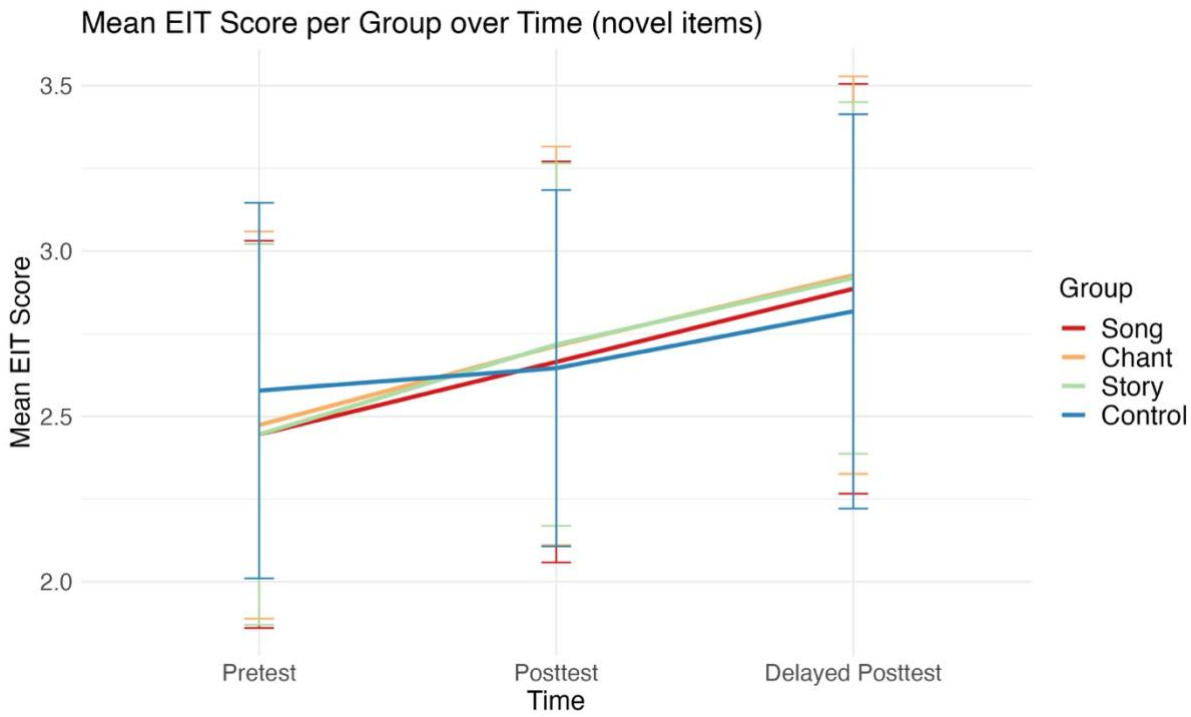
5.3.3.1 Descriptive statistics for RQ2c

There were 2248 responses to the eight stimuli containing novel items of vocabulary. The mean and standard deviation for each group at each time point are summarised in Table 5.6. Figure 5.5 plots the mean scores per group over time.

Table 5.6. Descriptive statistics for RQ2c

Group	Time	Mean	SD	SE	Lower 95% CI	Upper 95% CI
Song	Pretest	2.45	1.43	0.30	1.86	3.03
	Posttest	2.67	1.48	0.31	2.06	3.27
	Delayed posttest	2.89	1.52	0.32	2.27	3.51
Chant	Pretest	2.47	1.43	0.30	1.89	3.06
	Posttest	2.71	1.47	0.31	2.11	3.32
	Delayed posttest	2.92	1.47	0.31	2.33	3.53
Story	Pretest	2.45	1.41	0.29	1.87	3.02
	Posttest	2.72	1.34	0.28	2.17	3.27
	Delayed posttest	2.92	1.30	0.27	2.39	3.45
Control	Pretest	2.58	1.39	0.29	2.01	3.15
	Posttest	2.65	1.32	0.27	2.11	3.18
	Delayed posttest	2.82	1.46	0.30	2.22	3.41

Figure 5.5. Mean EIT score on novel items per group over time with 95% CI (RQ2c)



5.3.3.2 Model fit and diagnostics

The CLMM with two random effects failed to converge for the reduced data subset in RQ2c. R Studio returned an error message and only the estimates, with no SE, *z*-scores or *p*-values, both with Control as reference and Song as reference. To investigate potential multicollinearity among the predictors in the model, which could result in the model failing to converge, I calculated the variance inflation factor (VIF) for each predictor using the `vif()` function from the car package in R. The linear model specified was as follows:

```
lm(EIT_score ~ Song_VERSUS_Chant + Song_VERSUS_Story +
Song_VERSUS_Control, data = df)
```

The VIF values for the predictors were Song–Chant (1.53), Song–Story (1.52), and Song–Control (1.53). The VIF values were all less than five, which is typically the threshold for suggesting problematic multicollinearity between predictors (Kutner, 2005). Thus, while there is some correlation between these predictors, it did not appear to be multicollinearity that was

causing the CLMM to fail to converge. Having checked that the data itself did not contain any errors or blank values, I concluded that the full model was too complex to run with the reduced dataset and rebuilt the model with simpler terms. I began with the fixed factors only, then added the first random effect `(1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3|ID)`, and then a second random effect `(1|Sentence.ID)`.

The simplified model with two random effects (but no interaction in the random effects) converged and produced results in keeping with the trends observed in the descriptive statistics. I then ran a model comparison with the fixed factors model (CLM). The model with one random effect (CLMM) fitted the data much better than the CLM, as indicated by the highly significant likelihood ratio result (*LR.stat* = 717.78, *df* = 7, $p < .001$). Table 5.7 shows the model comparison statistics.

Table 5.7. Model comparison RQ2c

Model	**No. of Parameters**	**AIC**	**Log-Likelihood**
CLM	17	6481.0	-3223.5
CLMM	24	5777.3	-2864.6

5.3.3.3 Results of the CLMM for RQ2c(i)

A CLMM was used to test the interaction between contrasts between test times and contrasts between Control and experimental conditions. There was a main effect of time, with all groups improving their scores at posttest ($b = 0.41$, $SE = 0.1$, $z = 4.02$, $p < .001$) and delayed posttest ($b = 0.83$, $SE = 0.1$, $z = 7.95$, $p < .001$). There was a main effect of syllables, with scores decreasing in all groups as the EIT stimuli increased in length ($b = -1.57$, $SE = 0.24$, $z = -6.52$, $p < .001$).

RQ2c(i): Experimental conditions compared to control

All experimental conditions scored descriptively more highly than Control at posttest and delayed posttest. There were no significant interaction effects between contrasts between pretest and posttest and contrasts between Control and Song (b = 0.35, SE = 0.28, z = 1.23, p = .22, OR = 1.41), or Control and Chant (b = 0.37, SE = 0.28, z = 1.33, p = .18, OR = 1.45), or Control and Story (b = 0.54, SE = 0.28, z = 1.93, p = .054, OR = 1.71). There were no significant interactions between contrasts between pretest and delayed posttest and Control and Song (b = 0.497, SE = 0.29, z = 1.71, p = .086, OR = 1.64), or Control and Chant (b = 0.52, SE = 0.29, z = 1.78, p = .075, OR = 1.67). However, the interaction between the contrast between pretest and delayed posttest and contrast between Control and Story was statistically significant (b = 0.597, SE = 0.29, z = 2.08, p = .03720248). The odds ratio indicates that Story participants were 1.82 times more likely to score more highly than Control participants on the novel items at the delayed posttest.

RQ2c(ii): Relative effects between experimental conditions

For novel items, there were no significant interaction effects between contrasts between test times and between contrasts between Song and Chant, or Song and Story, as shown in Table 5.8.

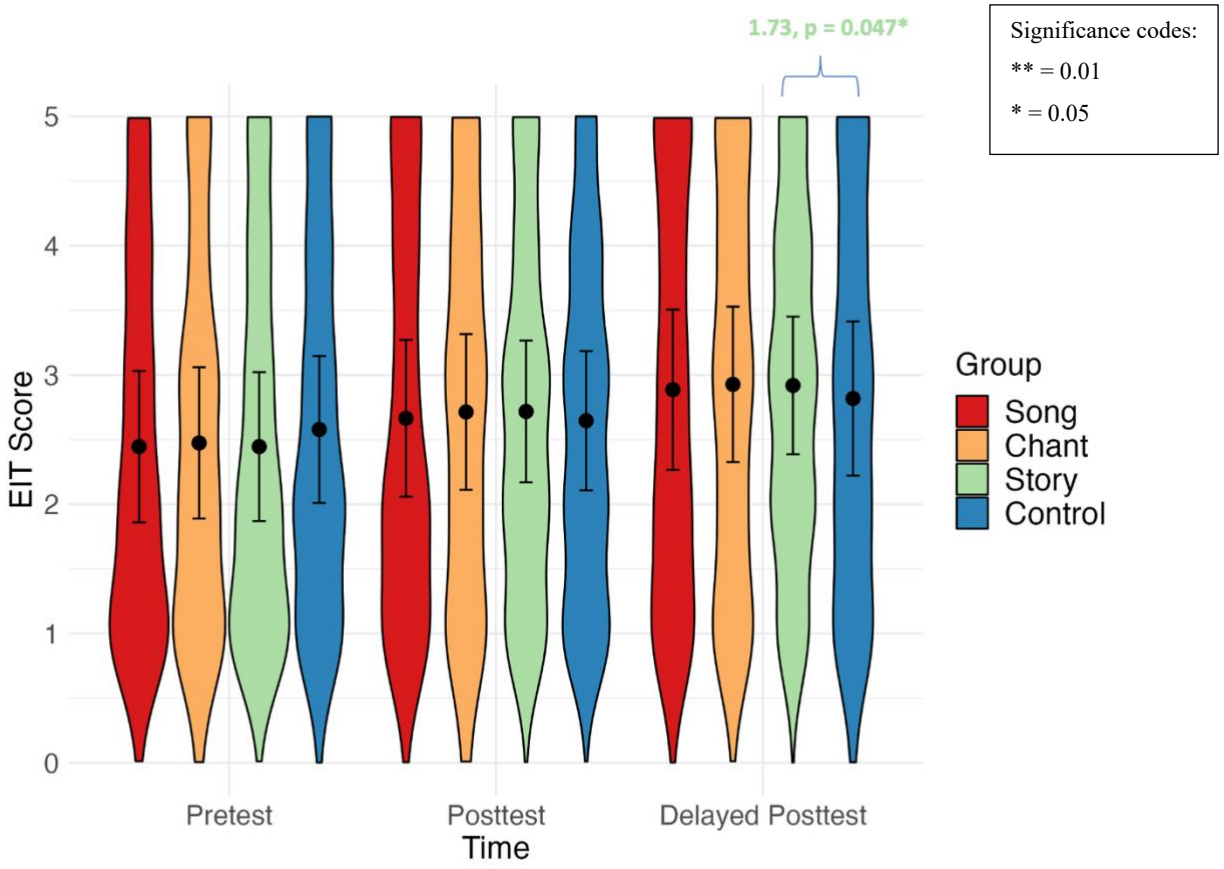
Table 5.8. Relative effects between experimental groups on familiar items RQ2c(ii)

Group/time	**Estimate**	**SE**	***z*-score**	***p*-value**	**Odds ratio**
Song and Chant * Posttest	0.03	0.29	0.09	0.93	1.03
Song and Chant * Delayed posttest	0.02	0.29	0.06	0.95	1.02
Song and Story * Posttest	0.19	0.29	0.66	0.51	1.21
Song and Story * Delayed posttest	0.10	0.29	0.34	0.73	1.11

The relative effects of the experimental conditions on participants' ability to reproduce novel items indicate very small differences between song, chant and story conditions at posttest and delayed posttest.

Figure 5.6's violin plot visualises the distribution of EIT scores (0 to 5) on novel items in each condition at each of the three test times, with the mean score and confidence intervals indicated with the dots and whiskers, and with the statistically significant odds ratio between Story and Control at delayed posttest superimposed.

Figure 5.6. Violin plot showing EIT scores on novel items, significant OR, group means and 95% confidence intervals (RQ2c)



In summary for RQ2c, there were no statistically significant interactions between the experimental conditions and test time. Only the interaction between the contrast between

pretest and delayed posttest and the contrast between Control and Story was statistically significant. The H_0 stated "there is no interaction between input condition and time on participants' performance on stimuli containing novel vocabulary items in the elicited imitation task." Based on the results presented here, we do not have evidence to reject the H_0 except in the case of Story compared to Control at delayed posttest. Experimental input condition does not appear to predict participants' ability to generalise their language knowledge to new contexts in the elicited imitation task. Whilst the mean differences between experimental groups are incredibly small, the CLMM analyses do not provide evidence *for the null.* We cannot state there is no interaction between experimental conditions and test time, only that there is no evidence of statistically significant interactions between test time and Song and Story or Song and Chant, or between test time and Song and Chant.

5.3.3.4 Results of Bayes factor analyses for RQ2c

I followed a similar process to RQ2a and RQ2b to address the problem of interpreting non-significant *p*-values by calculating Bayes factors. The relevant hypotheses were:

1) Participants in the song group are predicted to score more highly than participants in the chant group at a) posttest and b) delayed posttest.
2) Participants in the song group are predicted to score more highly than participants in the story group at a) posttest and b) delayed posttest.

Each of these is a one-sided prediction, thus I used a half-normal distribution to model the H_1, following Dienes (2008, 2021). For the estimate of the H_1 predicted mean difference I used the contrast between Story and Control, since this was the largest effect in the CLMM for RQ2c. This value was 0.536581 at posttest and 0.59729751 at delayed posttest.

For **hypothesis 1a**, there was no evidence either way of a difference between Song and Chant at posttest, M_{diff} = .04, b = 0.03, SE = 0.29, p = .93, $B_{HN(0, .27)}$ = 0.77, $RR_{B<1/3}[0.9:\infty]$.

For **hypothesis 1b**, there was no evidence either way of a difference between Song and Chant at delayed posttest, M_{diff} = .03, b = 0.02, SE = 0.29, p = .95, $B_{HN(0, .3)}$ = 0.72, $RR_{B<1/3}[0.9:\infty]$.

For **hypothesis 2a**, there was no evidence either way of a difference between Song and Story at posttest, M_{diff} = .05, b = 0.19, SE = 0.29, p = .51, $B_{HN(0, .27)}$ = 1.09, $RR_{B<1/3}[1.6:\infty]$.

For **hypothesis 2b**, there was no evidence either way of a difference between Song and Story at delayed posttest, M_{diff} = .03, b = 0.10, SE = 0.29, p = .73, $B_{HN(0, .3)}$ = 0.86, $RR_{B<1/3}[1.2:\infty]$.

To summarise, Bayes factor calculations testing the one-tailed prediction that Song is more likely to score higher than Chant or Story on novel items in the EIT at posttest or delayed posttest found that the data is insensitive. All B were between 1/3 and ≤3, indicating no evidence for either theory, thus no conclusion can be drawn beyond the absence of evidence. It is not possible to state that the experimental groups are equally likely to improve their performance on the EIT measure with novel items, just that (as found by the CLMM) they did not differ statistically significantly from one another and more data is required.

5.3.4 Research Question 2d

What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on their ability to correct a grammatical error in the *ne […] pas* negative word order in a subset of three previously encountered stimuli?

The 851 responses to the three EIT stimuli containing grammatical errors in the *ne [verb] pas* construction were coded 0 if the error was imitated, and 1 if the error was corrected by moving the *pas* to the correct place in the word order to produce *ne [verb] pas*. Six corrected responses out of a possible 851 were produced. Four corrections were made to item 4, *ce n'est pour pas* nous*, and two corrections were made to item 10, *il nous pas* mangera*.

Table 5.9 shows the six responses where a correction was made, and gives the transcription, EIT score and a commentary about the correction. Four responses (three for item 4 and one for item 10) were from Control participants, and two responses (one per item) were from Story participants. Three of the correct responses were given at the posttest, and three at the delayed posttest.

Table 5.9. Items where grammatical errors were corrected

Transcription of stimulus – item 4: Ce n'est pour pas* nous	**Group**	**Commentary**	**EIT score**	**Time**
Ce n'est pas pour nous	Control	Correct	5	Posttest
Je me pas no po	Story	Has moved the 'pas' to correct position but none of the other words correct	1	Posttest
Ce n'est voh pour nos	Control	Correct prosody but said voh not pas	3	Delayed posttest
Ce n'est pas comme	Control	Had the correct pas placement but incorrect sentence ending/ prosody	1	Delayed posttest
Transcription of stimulus – item 10: Il nous pas* mangera				
Il ne mange pas	Story	Unclear whether corrected grammar or switched 'era' from mangera with the 'pas'	1	Posttest
Ou nous pas mangiera pas	Control	Had both the correct and incorrect 'pas'!	2	Delayed posttest

Since there were so few correct responses, no inferential analyses were undertaken. In response to RQ2d, there is not enough data to make any inferential predictions but descriptively Control participants made the most corrections, followed by Story participants. No corrections were made by Song or Chant participants.

The H_0 stated that "there is no interaction between input condition and time on participants' ability to correct a grammatical error in the *ne […] pas* negative word order in the elicited imitation task." The data gathered here are insufficient to reject the H_0.

5.4 Summary of Phase 2 results

The gathered data and analyses conducted during Phase 2's intervention did not provide evidence to reject the null hypothesis (broadly that there is 'no interaction between input conditions and test time') on participants' overall performance on the EIT outcome measure

(RQ2a), their ability to reproduce language previously encountered in the input (RQ2b), their ability to generalise their language knowledge to novel items (RQ2c), or their ability to correct grammatical errors in the *ne [verb] pas* word order (RQ2d).

The non-significant interaction effects calculated using the CLMM do not provide evidence *for the null*, that there is no interaction between experimental conditions and test times. Nor, when tested using Bayes factor analyses, was there evidence to support the null hypothesis on RQ2a, RQ2b or RQ2c. Chapter 6 discusses these results in view of the existing literature, including Phase 1's systematic review. It considers the methodological limitations of Phase 2's intervention and suggests potential implications for research, practice and theory of the results presented in Chapter 5.

Chapter 6

Discussion

6.1 Introduction

This chapter reviews the context and aims of this research and summarises the literature review. Phase 1's systematic review findings are then summarised, and Phase 2's intervention study findings are interpreted and discussed. The implications for practice, research and theory of both phases are presented. The chapter closes by discussing the limitations of Phase 2, which should be considered when interpreting the findings or embarking on any replications of the study.

6.2 Context and aims: revisited

As a practising teacher, I encountered strong claims that songs constitute a (superlatively) effective tool for teaching YLLs and developing their FL skills. These assertions are echoed by generalist early years and primary school teachers as well as KS2 FL specialist teachers in England (Hamilton & Murphy, 2023), and YLL teachers from international contexts (Davanellos, 1999; Forster, 2006; Paquette & Rieg, 2008; Saricoban & Metin, 2000; Schoepp, 2001; Walker, 2006).

Songs have been one of the tools used for teaching FL since the song schools of the Middle Ages where oblates were taught the Latin mass through rote learning of sacred music (Kelly, 1976). As communicative competence and literacy replaced piety as a key focus of language education (Luscombe, 2004), songs continued to feature in manuals for FL teaching (Leach, 2005). Having stood the test of time and weathered multiple trends for how to approach teaching YLLs (see section 2.2.3), songs feature prominently in the 2014 statutory National Curriculum for primary FL (DfE, 2013a) and in earlier more extensive but non-statutory guidelines such as the *KS2 Framework for Languages* (DfES, 2005a), which was

used widely as a basis for primary FL teaching in England in the 2000s (Cable et al., 2012; Wade et al., 2009). Whilst primary FL was made a statutory requirement in England with the 2014 curriculum, the government failed to roll out any nationwide training or bank of resources to support primary schools to deliver the FL goal of helping pupils achieve "substantial progress" (DfE, 2013a). The KS2 Framework was seen as valuable by respondents in the DfE (2012) consultation on primary FL, and despite being removed from circulation by the coalition government, its legacy of using songs and rhythm for promoting oracy, literacy, intercultural understanding and metalinguistic awareness lives on in many of the commercial resources produced by languages specialists that are used in increasing numbers of primary schools (Collen & Duff, 2024).

Additionally, teachers report using songs in FL as a way of finding time for languages in the busy primary school curriculum (Tinsley & Doležal, 2018). Primary teachers who feel less confident in their language ability may use bought-in schemes of work that contain songs to provide authentic language content with correct pronunciation (Unopeia, 2022). There are blogs with frameworks for using songs in FL lessons at secondary level (Conti, 2015) and in primary schools (Míguez, 2017; Simpson, 2015). Some teaching approaches advocate for using traditional songs from the target culture (QCA, 2007) and others advocate making up the lyrics of songs from target items of vocabulary, sung to familiar English tunes, such as *London's Burning* (Forder et al., 2013), *Two Little Dickie Birds* (Kirsch, 2008) or *Old MacDonald* (Rumley, 1999). Songs are thus put to multiple uses in the primary FL classroom.

There appears to be an assumption that singing songs in the target language will automatically lead to communicative competence in contexts beyond the song being sung in a classroom (e.g., Kirsch, 2008). Yet, apart from drilling vocabulary in a more enjoyable way (Kirsch, 2008), the proposed mechanisms by which the linguistic proficiency needed to sing a song in class transfers to communicating in real life are under-specified. As suggested by the accounts of adults who know *Frère Jacques* but not what the words mean or how to use them

outside the context of the song itself, learners could be left with a fossilised internalised representation of a song if there is no 'exploitation' of the song for the purpose of explicitly learning language (Cameron, 2001; Conti, 2015). Perhaps when primary FL learners learn a song in the target language this leads to an improved sense of self-efficacy and a positive attitude towards learning languages with specialist teachers at secondary school (Rumley, 1999). However, the evidence reviewed in section 2.2 provided very little substantiation beyond the folk pedagogy of songs being a perpetual and natural part of FL teaching.

Thus, having established that songs are firmly part of teachers' repertoires for FL lessons in primary schools and seen as useful for multiple purposes, including linguistic outcomes as well as affective and classroom management purposes (Hamilton & Murphy, 2013; Smith, n.d.), the next section will recap the theoretical and observational evidence for using songs with YLLs.

6.2.1 Using songs: theoretical approaches

As well as songs being a 'folk pedagogy' handed down through generations of FL teachers, some of the theoretical basis for research investigating songs and language outcomes is also based on 'folk *theory'* (Bruner, 1996, emphasis added). In his theoretical review, Engh (2013) proposed several avenues of theoretical motivation for using songs to teach languages, but there was limited empirical support for the theories themselves. For example, research on musical learning styles (see 2.3.1.2) is not a sound foundation for experimental work to build upon because, although often invoked in support of using songs for language learning (e.g., Fonseca-Mora, 2000; Engh, 2013), there are scant rigorous or replicable findings linking learner preferences to pedagogy (Coffield et al., 2004). Another highly cited theory is the 'song stuck in my head' phenomenon (SSIMH; Murphey, 1990), which proposes that songs may be beneficial for language learning because they sometimes echo in learners' heads, producing a practice effect. Itself based on unfalsifiable and unsubstantiated theories (particularly Krashen's 1983 'din' hypothesis and Chomsky's 1965 'language acquisition

device'), Murphey's (1990) SSIMH arose from a pilot survey of 49 EFL students who agreed when prompted that they too experienced involuntary mental song rehearsal. Despite being an inadequate evidential basis for claims about songs' mnemonic and ensuing linguistic benefits for FL learners, SSIMH continues to be cited, in much the same vein as the musical learning styles literature, particularly in teacher-produced or oriented publications (Davanellos, 1990; Degrave, 2019; Fonseca-Mora, 2000; Thain, 2010).

Providing a more reliable and valid source of theoretical motivation is the 'prosodic bootstrapping hypothesis' whereby infants learn how to map phonological prosodic contours from the input onto meaningful linguistic units such as utterances or phrases, bootstrapping the lexis and syntax of their L1 (Gleitman & Wanner, 1982; Morgan & Demuth, 1996). As a well-evidenced theory of L1 acquisition, prosodic bootstrapping helps to explain how infants exploit phonological cues to begin segmenting and categorising the speech stream, using species- and domain-general learning mechanisms (Hauser et al., 2001; Saffran et al., 1999). Whilst there are still open questions about which prosodic cues are exploited for which languages and at different stages of development, this "powerful heuristic" (Gervain et al., 2020: 573) in L1 acquisition seems to remain online throughout childhood and into speech processing in adulthood (Streeter, 1978).

In L2 acquisition, there is some debate about whether an established L1 prosodic system competes with or complements an emerging L2 system (Goad & White, 2004, 2019; Jongman & Tremblay, 2020). It seems, however, that with enough exposure, adults can exploit target language prosodic cues to learn novel lexis and word order (Saksida et al., 2021). Two studies by Campfield and Murphy (2013; 2014) suggest that prosodic bootstrapping processes extend to YLLs learning EFL in Polish classrooms when the input is made more rhythmically salient through the use of nursery rhymes.

6.2.2 Evidence from experiments

Given the strength of practitioner belief in the linguistic benefits of using songs for teaching YLLs in FL classroom contexts, there appeared to be scant robust empirical evidence supporting the use of songs to achieve linguistic development with YLLs according to existing narrative and semi-systematic reviews of empirical work on the topic (Davis, 2017; Degrave, 2019; Sposet, 2008; Werner, 2020). The overall weight of collective evidence presented in these reviews was unclear, however, since none of them assessed the methodological quality of their included works, and incomplete reporting leaves questions about the exhaustiveness of their search strategies. It was therefore not possible to draw firm or meaningful conclusions about the substantive linguistic effects of using songs to teach YLLs from existing reviews. This thesis therefore first aimed to provide a thorough assessment of the state of our collective knowledge about the causal link between songs and linguistic outcomes, and then (in response to the findings of the systematic review in Phase 1), to design a methodologically rigorous study that compared teaching FL through songs to teaching FL through chants or stories, in Phase 2.

Recall that I have deliberately taken a 'strong' position on what constitutes 'evidence' of songs' effectiveness in FL learning (see section 1.2.1) due to my perception that progress in this research field is hampered by confirmation bias and attempts to derive causal claims from study designs that are neither appropriate for nor sufficiently robust to detect causal relationships. Multiple research designs are potentially useful in pursuit of research evidence for purposes such as determining why or how something works, but for questions of 'what works', there is an epistemological tradition in the social sciences clearly laid out by Campbell (1957) and elaborated upon by Campbell and colleagues (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002) and others (Connolly et al., 2017; Gorard, 2003, 2013; Slavin, 1986). This tradition entails making formal comparisons, ideally using a randomised trial design, and if not, using as robust a design as possible given the context (see section 1.2.1

for elaboration). Taking this approach does not discount the valuable contextual evidence provided by studies asking different questions, and using different, non-experimental, designs (e.g., Geisler, 2008; Kamimski, 2016), as long as their (limited) capacity to reliably address casual questions is acknowledged. However, before evaluating the mechanisms through which songs might be best employed for motivating children to develop their language skills, it is first necessary to establish whether we know that songs can be attributed with any positive causal effect (other than an assumed one) on YLLs' linguistic outcomes relative to other teaching methods. Hence my study focused on testing the claim that songs 'work' for teaching languages, rather than exploring how or why they might work.

6.2.3 Summary of the literature review and next steps

Based on work summarised in the literature review, I identified a gap between the practice-based, experiential evidence of teachers and teaching materials and the research literature coming from experimental studies of learners in FL classrooms. Where using songs for teaching FL is presented as not only a valued but a valuable practice for the purpose of achieving children's linguistic outcomes in the former, there appeared to be scant evidence from the latter to support conclusions about the causal link between using songs to teach FL and linguistic outcomes. There also appeared to be some 'folk theory' from learning styles, musical intelligence and 'song stuck in my head' surveys circulating in the field that did little to motivate well-evidenced approaches to researching the causal links between using songs and linguistic outcomes. Two classroom studies (Campfield & Murphy, 2013, 2014) drew on the prosodic bootstrapping hypothesis, which provided a more substantial foundation to build research upon, but these studies investigated nursery rhymes compared to stories, rather than songs. The question of what the effects of using songs in FL lessons on the substantive linguistic outcomes of YLLs are appeared to be answered neither satisfactorily nor conclusively.

However, the narrative approach I took to assembling the literature review could have left out important evidence from empirical research that would inform teachers about whether choosing songs as a form of FL input in classroom activities has a demonstrable effect on linguistic outcomes. I therefore prepared the systematic review in Phase 1 of this study to locate, synthesise and appraise research that directly investigated the effects of using songs, chants and nursery rhymes (howsoever that is operationalised as a whole-class FL activity) on the linguistic outcomes of children aged 2 to 18 years old in formal educational settings. The full discussion of findings from this systematic review is presented in Chapter 3 (section 3.6), with a summary presented in section 6.3 below.

6.3 Phase 1: Systematic review of intervention studies

6.3.1 Summary of findings

The systematic review sought intervention studies published in any language on any date conducted in preschool, primary or secondary schools with children aged 2–18 that assessed the effects of using songs, chanting or nursery rhymes on linguistic outcomes. English, French, German and Spanish keyword searches of education, psychology and thesis databases located 2868 records. After deduplication, screening of abstracts and the 89 subsequently located full texts, 60 records were found to contain reports of intervention studies meeting the inclusion criteria, 50 of which were published in English. These include studies published between 1978 and 2021 from 23 countries that assess the relationship between using songs in the classroom and substantive linguistic outcomes of learners. Studies focus primarily on vocabulary outcomes (35 studies include vocabulary measures), followed by grammar, speaking, reading, listening and writing outcomes. Over half (34) studies were conducted with primary school learners, with the remaining 26 studies split equally between preschool and secondary school contexts. Only one study conducted in the UK (Jarvis, 2013) involved learners in an Early Years Foundation Stage classroom, but since the age of participants was not reported it could be that the learners were aged anywhere between 3 to 5 years old (since

the EYFS finishes at Reception, the first year of primary school). It is possible, then, that no relevant studies have been conducted in UK primary schools.

Heterogeneity of methods and demographics precluded a meta-analysis of statistical analyses, but a narrative synthesis and risk of bias assessment found that the assembled literature did not provide convincing and reliable evidence of songs' direct benefits for YLLs' linguistic outcomes, despite positive claims made in a majority of studies. Of the 60 included studies, three received 'strong', fourteen 'moderate', and 43 'limited' global weight of evidence ratings. Given the overall scarcity of reliable and conclusive evidence from studies conducted with experimental rigour, the systematic review found very limited evidence to support the strong and long-standing belief held by many YLL teachers that songs are an effective FL pedagogy, with demonstrable effects on children's language development in FL classrooms. Indeed, while much of the assembled literature made positive causal claims about the relationship between singing songs and linguistic outcomes, a majority were not appropriately designed to support these claims.

6.3.2 Implications from the systematic review

With such a vanishingly small body of reliable research conducted in the field overall, there was almost no reliable evidence that UK primary FL teachers contemplating using songs as a tool for language teaching could draw upon to inform their pedagogical choices. This finding was in opposition to the strong support for using songs for FL teaching seen in documents such as the *KS2 Framework for Languages* (DfES, 2005a) and England's current National Curriculum for FL (DfE, 2013a).

This doctoral study's first contribution to the field was thus to have systematically searched, located, synthesised and dispassionately appraised evidence from intervention studies investigating the substantive linguistic outcomes of YLLs in second or foreign language classrooms where songs were used as a whole-class activity, and to find that our collective knowledge on this topic, at best, lacks substance. Coupled with the absence of any

rigorous studies (or possibly any studies at all) conducted in UK primary school settings, there was therefore a warrant to conduct a well-designed study in a UK primary school context. This proposed study formed the foundation for Phase 2's original contribution to address the absence of evidence to support English primary FL teachers' pedagogical use of songs detected in Phase 1's systematic review.

6.4 Phase 2: Intervention study

Following Phase 1's systematic review, this thesis' second original contribution to the field is a randomised controlled trial comparing the relative effects of using songs, chants or stories as the source of FL input on primary school beginner learners' performance in an elicited imitation task. The three experimental input methods were compared to each other, and to a control group who received an equivalent length of exposure to French input, but using the 'business as usual' primary language materials provided by the schools rather than input through songs, chants or stories.

The research design built on findings from two studies of Polish primary school EFL learners by Campfield and Murphy (2013, 2014). In particular, this study's design followed Campfield and Murphy's 2014 study which was an RCT comparing a rhythmically salient English input condition (nursery rhymes) to a prose input condition (stories), with a control group who received their usual EFL lessons. Campfield and Murphy (2014) found that the rhythmically salient input group achieved higher performance in an elicited imitation outcome task containing 22 stimuli ranging in length from five to nine syllables. The current study adds a song condition to the experimental input to ascertain whether rhythm plus melody confers any additional benefit.

Additionally, by randomly allocating participants to conditions and having an active control condition, this thesis extends the work of Davis and Fan (2016) which was a single group pre/post design investigating Chinese preschool learners' productive English vocabulary with three input conditions: song, chant (choral repetition) and 'no presentation'

control. Randomly allocating and giving the control group a valid comparison teaching method provides a more experimentally reliable and ecologically valid control comparison, since teachers are more interested in how to present the FL input, not whether to present it at all (as Davis and Fan's 'no presentation' method implies).

Phase 2's intervention also uses a mixed-effects statistical analysis method (the CLMM) to account for the hierarchically structured data derived from repeated-measures experimental designs. It thereby adopts more robust statistical methods than either of the Campfield and Murphy (2014) or Davis and Fan (2016) studies, which used MANCOVA and ANOVA respectively. Using mixed-effects models helps to account for individual differences across time points and, unlike MANCOVA and ANOVA methods, does not have the assumption of independent data points which are violated by repeated-measures designs. Combined, the use of an RCT design with more robust statistical methods helps to minimise the potential bias of previous similar studies and more effectively isolate the relative effects of input condition on the outcome measure. The following sections discuss the findings for each of Phase 2's research sub-questions.

6.5 Findings of the intervention study

The research question for Phase 2 asked:

RQ2: What are the effects of presenting and rehearsing linguistic input in the form of songs, chants or stories compared to:

i) a business-as-usual control condition and

ii) each other

on beginner primary school French learners' performance in an elicited imitation task (EIT) on:

a) all 22 stimuli?

b) a subset of fourteen previously encountered stimuli?

c) a subset of eight previously encountered stimuli containing novel vocabulary items?

d) their ability to correct a grammatical error in the *ne […] pas* negative word order in a subset of three previously encountered stimuli?

This investigation entailed two 'contrasts' being executed in the CLMM statistical analysis for part i and part ii of RQ2 a, b and c: one with Control as reference group for part i, and one with Song as reference group for part ii. RQ2a included analysis of the responses to all 22 of the EIT stimuli at all three time points (pretest, posttest and delayed posttest). RQ2b and c each took a respective subset of stimuli into account. For RQ2d, a subset of 851 potentially relevant responses was analysed. The following sections interpret and discuss the findings for each subquestion, and then address overall implications for practice and research of this intervention study.

6.5.1 Findings for RQ2a: performance on all 22 stimuli

Regarding part i of RQ2a, there was a statistically significant interaction between the contrasts between Control and Song conditions and the contrasts between pretest and posttest, and the contrasts between Control and all three experimental conditions and the contrasts between pretest and delayed posttest. Regarding part ii, the study did not detect any statistically significant interaction effects between the contrasts between Song and Chant or Song and Story and the contrasts between pretest and posttest, or pretest and delayed posttest. Since a non-significant *p*-value in a null hypothesis test cannot provide evidence of no effect (i.e., evidence for the null), Bayes factors were computed to test the H_0 that there was no difference between Song and Chant or Song and Story conditions at posttest and delayed posttest. However, the data were insensitive: more data would be required to provide anything more than ambiguous support for the H_0 in each case.

The data do not provide evidence that the song condition participants outperform participants in the other two input conditions on the EIT outcome. Song, Chant and Story participants all made similar progress across time points, hence the null hypothesis that there

is no interaction between group and time could not be rejected. Nor, however, do the data provide evidence for the hypothesis that there is no difference between the Song and alternative experimental input conditions across time points. We cannot therefore conclude that the input conditions are the same, just that there is no evidence that they are different.

We can deduce from the statistically significant interactions between the contrast between Control and Song and the contrast between pretest and posttest, and the contrasts between Control and all three experimental groups and the contrast between pretest and delayed posttest that, on average, experimental group participants were able to perceive the EIT stimuli within the songs, chants or story input forms and rehearse the content. This enabled them to perform more highly on the EIT than the Control participants, who were not presented the EIT stimuli in their input. Hence it is possible to conclude that presenting the input in the form of songs, chants or stories was similarly, if not equally, effective, and that learning did take place, when compared to the 'business-as-usual' control condition which entailed participants receiving the same length of French exposure, without inclusion of the EIT stimuli in their input.

If there had been no active control group who received a comparable amount of French from the same teacher as the experimental groups, it would be impossible to differentiate between a non-significant interaction effect or a failure to detect any experimental effects at all. As it stands, it is possible to conclude that the data only show main effects of group and time (i.e., all groups made progress over time to some extent) and interaction effects between contrasts between Control and experimental groups and contrasts between time points, but no interaction effects between the contrasts between the three experimental groups and the contrasts between timepoints. The 95% confidence intervals all overlap; hence the only robust conclusion is that the study is underpowered to detect effects of such a small magnitude if they exist. Since participants were individually randomly allocated to conditions, which increases statistical power to detect an effect, if a large effect were to

occur, it would surely have been detectable. Thus, despite not supporting a conclusion of exactly equal (or no) difference, the data suggest any relative effects of input condition are, all things being equal, incredibly small.

Since a similar pattern emerges from the subset analyses, I will present these first before discussing overall interpretations for RQ2a, b and c together.

6.5.2 Findings for RQ2b: performance on a subset of fourteen previously encountered stimuli

A subset of fourteen EIT stimuli containing no changes from the input for the three experimental conditions were analysed for RQ2b. The findings present a similar pattern to RQ2a where the overall EIT responses are considered. For RQ2bi, there was only a statistically significant interaction between the contrasts between pretest and delayed posttest and the contrasts between Control and Song or Control and Chant (but not Control and Story, nor any contrasts between pretest and posttest). For RQ2bii, in this subset of 3934 responses, there were no statistically significant interactions between contrasts between Song and Chant or Song and Story and contrasts between test times. Despite the vanishingly small mean differences between experimental groups, Bayes factor analyses did not find that the evidence supports a null hypothesis of no difference between Song and Chant or Song and Story at posttest or delayed posttest. Hence the data appear to be insensitive. More data is required to draw anything other than ambiguous conclusions from the Bayes factor analyses.

The findings from this subset analysis of items where experimental group participants had already encountered all of the stimuli during presentation and rehearsal of the input suggest that input condition does not explain variance in the outcome. Isolating only the familiar items does not change the picture from RQ2a where all EIT stimuli responses were analysed. We have no evidence of relative difference in performance over time between experimental groups, and no evidence of exactly equal performance over time either. The

following section looks at the findings for EIT responses where a novel item of vocabulary is included.

6.5.3 Findings for RQ2c: performance on a subset of eight previously encountered stimuli containing novel vocabulary items

Investigating RQ2c involved analysing the 2248 responses given over the three time points to eight EIT stimuli where a previously encountered (familiar) item of vocabulary from the input was substituted for a novel item of the same length and prosodic structure. The rest of the EIT stimulus in each case remained as it had been encountered in the input. For example, the phrase *Mon petit lapin* (as it appeared in the input) was altered to *Mon petit bouquin* in the EIT stimulus. Responses to these items revealed that the pattern of performance from RQ2a and RQ2b remained stable for stimuli containing novel vocabulary substitutions.

Regarding RQ2ci, only the interaction between the contrast between pretest and delayed posttest and the contrast between Control and Story was statistically significant. No other interactions between Control and Song or Control and Chant and contrasts between pretest and either posttest was significant. For RQ2cii, there were no significant interaction effects between contrasts between test times and between contrasts between Song and Chant, or Song and Story. There was no evidence to reject the null hypothesis. Equally inconclusively, Bayes factor calculations testing the one-tailed prediction that Song is more likely to score more highly than Chant or Story on novel items in the EIT at posttest or delayed posttest found that the data are insensitive. There was ambiguous evidence of no difference (the H_0) in performance between these contrasts.

These non-significant results suggest that there is a vanishingly small difference between input conditions on the EIT stimuli containing novel items of vocabulary: all groups improved their performance over time, but no groups performed relatively more highly other than Story compared to Control at delayed posttest. Participants in all groups thus appeared to generalise their learning from the input to novel language contexts similarly successfully.

6.5.4 Interpretation of findings for RQ2a, b and c

The data gathered from the EIT responses suggest that there is very little to differentiate the EIT performance of beginner French learners who are presented input in the form of songs, chants or stories and given time to rehearse the input in these forms orally. Compared to a control group who received the same amount of class time learning French through listening and speaking, with similar visual orthographic support on the screen, but no rehearsal of the EIT stimuli (other than the occasional vocabulary item such as *lapin*), the three experimental groups performed relatively similarly. Participants were randomly allocated and groups were homogenous on the screening variables, hence this finding cannot be explained by any pre-existing factors that might bias their performance such as non-verbal IQ, English or French vocabulary knowledge, or rhythmic ability. Furthermore, mixed effects statistical analyses account for the hierarchical structure of the repeated-measures data (i.e., observations nested in individuals in time points). The estimates produced by the CLMM thus can be considered robust.

The findings from RQ2a, b and c are now compared in more detail to the findings of Campfield and Murphy (2014), the study upon which this one is based. To recap, Campfield and Murphy (2014) randomly allocated 80 Polish EFL primary school learners (mean age 8;4) to rhythm, prose and control groups. The rhythm group received English input through 20 nursery rhymes. The prose group received English input through a twelve-part story composed from the nursery rhyme words. The control group received their scheduled English lessons with no additional input and were tested later to compensate for the additional exposure in the experimental groups. The intervention lasted for twelve hours, which is substantially longer than the four hours of intervention time in this study. A MANCOVA, conducted on the categorical EIT pretest response ratings (0–5, the same as in this study) with screening variables included as covariates, revealed no between-groups differences in the EIT ratings received at pretest. Posttest response categories (adjusted for the effect of pretest

responses) revealed statistically significant differences between groups in terms of their response ratings ($F(12) = 4.46$, $p < .01$). Post-hoc pairwise comparisons revealed that statistically significantly more exact imitations (the highest score of 5 on the scale) were made by the rhythm group compared to the prose ($MD = 1.40$, $SE = .55$, $p < .05$) and control groups ($MD = 2.79$, $SE = .56$, $p < .01$), and the prose group compared to the control group ($MD = 1.40$, $SE = .55$, $p < .05$). A further ANOVA revealed statistically significant differences in the improvement scores between groups ($F(2,1691) = 68.67$, $p < .01$, with a medium effect size of *omega* = .31), with statistically significant planned contrasts between rhythm and prose ($t(1153.93) = 3.79$, $p < .01$, with a small effect size of $r = .11$), and both experimental groups and control ($t(1422.12) = 12.36$, $p < .01$, with a medium effect size of $r = .31$). The rhythm group thus appeared to make greater improvement from pretest to posttest. Kruskal-Wallis between-groups tests of the EIT response data by item length (the number of syllables) were significant at all lengths above four syllables (the shortest length tested). Mann-Whitney tests revealed that the rhythm group performed slightly better on some categories of longer sentences (six and eight syllables) than the prose group but with very small effect sizes (six syllables: $U = 6979$, $r = .18$; eight syllables: $U = 7277$, $r = .15$), and both outperformed the control group on sentences above four syllables long.

Overall, the differences between both experimental groups and control are generally statistically significant with medium effect sizes, whereas the difference between rhythm and prose are much smaller with small effect sizes. Indeed, performance in the rhythm and prose groups is "remarkably similar" (Campfield and Murphy, 2014: 216). Campfield and Murphy conclude that there are benefits to exposing children to longer stretches of language as found in stories and poems, and that the rhythm-salient condition appeared to improve participants' fluency and accuracy, albeit with small effect sizes and potentially attributing this finding to the context of individual EIT stimuli in the input rather than the rhythmic salience of the

groups. It could be that the rhythm group were exposed to more repetitions of some sentences than the prose group and vice versa.

Despite, on first glance, appearing to have detected different effects to this earlier study, on closer inspection there is a striking similarity in the findings of my study compared to Campfield and Murphy (2014). Firstly, the experimental groups generally outperform the control group. In both studies, the control group was not exposed to the content of the EIT outcome measure in their input: it is an experimental control that is present as a means of determining whether the intervention conditions have any effect overall compared to a baseline (as opposed to relative to each other). One conclusion in terms of methodology, then, is that our similar experimental design has achieved a similar effect. Whilst the control group improved over time, they did not do so at the rate of the experimental groups, presumably because they were not exposed to the EIT stimuli in their input.

In my study, notably, the length of time spent learning French (240 minutes) was the same for all groups, with the control group receiving the same treatment of being taken out of class and given a French lesson, rather than continuing with their usual FL lessons, which may have confounded Campfield and Murphy's results due to the 'special treatment' of the experimental groups. All of my groups received their posttest and delayed posttests within the same time frame, whereas there was a delay for Campfield and Murphy to allow for the control group to continue with EFL lessons to catch up with the intervention groups' time on task. This differential treatment does not appear to change the pattern of results between the two studies, but the equality of treatment enabled a fairer comparison to be made in my study, with no confounding factor of time delay about when the control group received their posttest. The key point from both studies is that, by not exposing the control group to the EIT stimuli but giving them an equivalent amount of FL exposure, a comparison of relative effects between the experimental groups can be made more effectively. Plus, unlike in Davis and Fan (2016) where the experimental taught (song/chant) conditions are compared to 'no teaching', it

is an ecologically valid comparison and thus potentially more useful for teachers to base their practice decisions on.

Secondly, the experimental groups in the current study and in Campfield and Murphy (2014) score very similarly on the EIT outcome measure relative to each other. There is a slight indication that Campfield and Murphy's rhythm group scored more exact imitations, the highest score on the EIT rating scale, relative to the prose group. Even so, their effect size is $r = .11$ for the contrast between improvement scores for rhythm and prose groups, which is a very small effect size. Therefore, when examining the provenance of Campfield and Murphy's (2014) significant finding of a relative effect between experimental conditions in more detail, it seems to all but disappear: for most of the statistical tests carried out, the rhythm and prose groups perform similarly compared to each other (see further statistical tests reported in Campfield, 2010, such as the discriminant analysis where rhythm and prose are not differentiated from each other, but are again differentiated from the control). It is only for the exact imitation EIT rating category and for the overall improvement score (which has a very small effect size) that the rhythm and prose groups are statistically significantly different to each other. My study did not find any statistically significant interaction effects between experimental groups, which on first glance appeared to be an apparent failure to replicate the statistically significant findings of Campfield and Murphy (2014) in an English FL classroom with similar aged learners. However, since there is only the smallest indication that the rhythm and prose groups differ in Campfield and Murphy (2014), and the MANCOVA does not actually test for the interaction effects between group and time, there may therefore not be much to differentiate between the two studies' findings after all.

What difference there is between the studies' findings could be partly attributed to the different statistical modelling approaches taken. The current study used a mixed effects model specifically for ordinal outcome data (CLMM) and tested for a Group*Time interaction effect. Campfield and Murphy (2014) used a MANCOVA which assumes a multivariate normal

distribution and continuous outcome, thus ignoring the ordinal nature of the outcome since it cannot be assumed that steps between the points on the EIT scale are equal intervals. The CLMM takes into account data clustering and does not have the assumption of independence as a MANCOVA or ANOVA does. Violating the assumption of independence inflates the Type 1 error rate in the *F*-statistic. Also, there is only one posttest in Campfield and Murphy (2014) thus only one change score per participant, whereas the current study's CLMM has delayed posttest data taken into account. This additional data provides an increase of statistical power, as well as providing longer-term indications of performance across the groups. Taken together, the potential for Type 1 error in Campfield and Murphy's (2014) MANCOVA and the small effect sizes indicate that the statistically significant difference detected between rhythm and prose may not be an entirely reliable finding and, if the data were reanalysed with a more robust model such as a CLMM, there may be no cause for rejecting the null hypothesis that there is no difference between groups.

As well as differing statistical approaches, differences in the target languages (English for Campfield and Murphy, 2014, and French in this study) provide another avenue for variation in the results. English rhythm follows an 'intensity' prosodic pattern and French a 'durational' pattern (Hayes, 1995). Since they had not had any prior French exposure before engaging with the interventions (see section 2.3.1.3.5), the English beginner learners of French in this study may not yet have a well-developed ear for the syllable-timed nature of French rhythm, where only the final syllables of lines are emphasised to align with natural speech rhythms (Dell & Halle, 2009; Kiparsky, 2020). Indeed, the learners in Campfield and Murphy (2014) already had one year (approximately 54 hours) of English lessons. They may have thus been more attuned to the prosody of English, which could confer an advantage in parsing the speech stream during the EIT. Or it could be the case that there is less rhythmically-salient differentiation between the story condition and chant/song conditions in this study because French rhythm contains a final-syllable emphasis in each line of verse, and

the story was read aloud and 'back-chained' which broke up the sentences into shorter sections while participants rehearsed the input. Although each sentence was read as a whole at the start of each session and again at the end, participants would have been hearing shorter chunks of each sentence whilst back-chaining it which might have made the story input more similar to the rhythmically-salient conditions (albeit without the melody).

Davis and Fan (2016) also found no evidence of a difference between song and chant conditions in their study. Considering that their chant condition entailed whole-class rehearsal of phrases, rather than lines or phrases taken from prosodically salient rhymes or poems, and their analysis only contained five items per condition, it is challenging to conclude that our studies' findings are aligned. The chant input conditions are too different in the two studies to draw a comparison. Furthermore, Davis and Fan (2016) could not robustly conclude that they have evidence of no difference since they did not gather evidence *for* their null hypothesis. Despite calculating Bayes factors, my study has not found evidence for the null hypothesis that there is no difference in performance between Song and Chant. More data is required before declaring that performance was equal in these conditions. Both studies are inconclusive on this point for different reasons, but it can at least be claimed that the current study has higher statistical power having randomly allocated individuals to conditions in a more robust experimental design for detecting between-subjects effects, and a more reliable statistical analysis that does not violate the assumption of independence.

Overall, then, it can be concluded that the current study data indicate that similar outcomes on an EIT are produced when beginner French learners (aged 7–8 years) in two English primary schools are presented with French input in the form of songs, chants or stories that they listen to and rehearse orally over the course of three weeks of lessons. The intervention did not indicate that presenting and rehearsing input in the form of songs was more effective than presenting input in the form of chants or stories. There was no evidence of relative differences in performance over time between experimental input conditions, but

equally, no evidence of equality between them. These findings align quite closely with those of Campfield and Murphy (2014) whose study produced remarkably similar data for Polish EFL learners in rhythmically salient or prose conditions: there was very little to differentiate between these conditions statistically, other than a small effect of increased improvement in the rhythm group and higher numbers of exact imitations. Their statistically significant findings could, however, be attributed to processes that leave the analyses prone to Type 1 errors. Davis and Fan (2016) also found no statistically significant differences between song and chant conditions in their within-subjects study. The current study, using more robust statistical methods, found no relative advantage for one experimental input condition over another. It cannot thus be said that songs are more facilitative of French acquisition for beginner learners, but that all three input conditions (and indeed the control group) made progress over time, with the only statistically significant interactions occurring between contrasts between Control and experimental groups over time, but not between experimental groups over time.

6.5.5. Findings for RQ2d

Question RQ2d asked about participants' ability to correct a grammatical error in the *ne […] pas* negative word order in a subset of three previously encountered stimuli. Across the three time points, there were 851 responses with potential 'corrections' available. Only six responses made any form of correction to the negative word order. Only one of these six, from a Control participant at the posttest, scored 5 for an exact imitation of the stimulus with corrected negative: *Ce n'est pas pour nous*. The remaining corrections (three for the same stimulus, and two for *Il nous pas* mangera*) did not score above 3 and some appeared to be accidental or the word order correction was the only correct part of the imitation, with the prosody and other lexical items being incorrect (as in the response *Je me pas no po* for *Ce n'est pas pour nous*). There were no corrections for the third sentence, item 18: *Pendant que le loup n'y pas* est*. This stimulus was perhaps too long (eight syllables) and beyond the capability of the

beginner learners. It appears that the participants had not received enough input to infer the correct negative French word order from their 240 minutes of input.

Alternatively, it could be that the participants were simply following my instructions to repeat what they heard on the recording. Perhaps they did notice the word order violation but did not correct it. However, it seems unlikely that after 240 minutes of input they would have an explicit understanding of the negative word order, notice a violation of it in the stimuli, and then repress a correct imitation or fail to mention it to me. The most likely interpretation is that the participants did not yet have a mental representation of the French negative that would enable them to process the input and correct the error.

This finding therefore provides a small amount of novel information about the developing interlanguage of beginner French learners. After three weeks of regular French lessons containing the negative clause, they had not implicitly (since there was no explicit instruction) learned the negative word order. None of the experimental input conditions conferred an advantage in this aspect of the study, as far as the quantitative data indicate. Indeed, the only wholly correct response came from a Control participant. There were no negative items presented in the Control materials, hence no opportunity to learn the negative word order from the input. This single correct response therefore seems to have been produced entirely by chance, reinforcing the conclusion that 240 minutes of input is insufficient to learn this grammatical feature.

6.6 Implications for practice, research and theory

There is a remarkable mismatch between the strength of belief in the causal link between using songs for FL teaching and children's linguistic outcomes, and the strength of empirical evidence that would support teachers' choice of this valued teaching practice for meeting linguistic outcomes. Indeed, the evidence base is notable for its lack of substance, which is surprising because anecdotal accounts and teaching manuals for primary FL describe songs as superlatively effective FL teaching material, particularly with young learners. There are

numerous reasons why teachers might choose to use songs in an FL lesson other than for achieving a specific linguistic outcome. Songs are perceived as being motivational (DfE, 2013b), making repetition more fun (Kirsch, 2008), providing relevant intercultural content (Light Bulb Languages, 2025), and as being a source of authentic language (Unopeia, 2022). Songs are woven into early educational provision (Hamilton & Murphy, 2023) and it feels 'natural' (Léopold et al., 1969) and indeed 'sensible' (Light Bulb Languages, 2025) to do songs with YLLs. Indeed, it would feel unnatural, perhaps, not to introduce children to the songs of the target culture. Who ever heard of *not* learning *Frère Jacques* at primary school?

Rather than investigating the myriad ways that teachers might choose to use songs for FL teaching, or the routes through which songs might provide motivation for YLLs to learn languages in a declining context (Macaro, 2008), this study simply addressed the question of whether using songs in FL teaching with YLLs can be linked to substantive linguistic outcomes. In other words, do primary school children learn elements of the target language when presented through songs? And, if they do, can this learning be generalised to novel lexical items, beyond the specific items heard in the input? It seemed that this question of basic science was not addressed effectively in the literature and needed to be established before investigating the mechanisms (how and why) of songs' putative links with language learning outcomes.

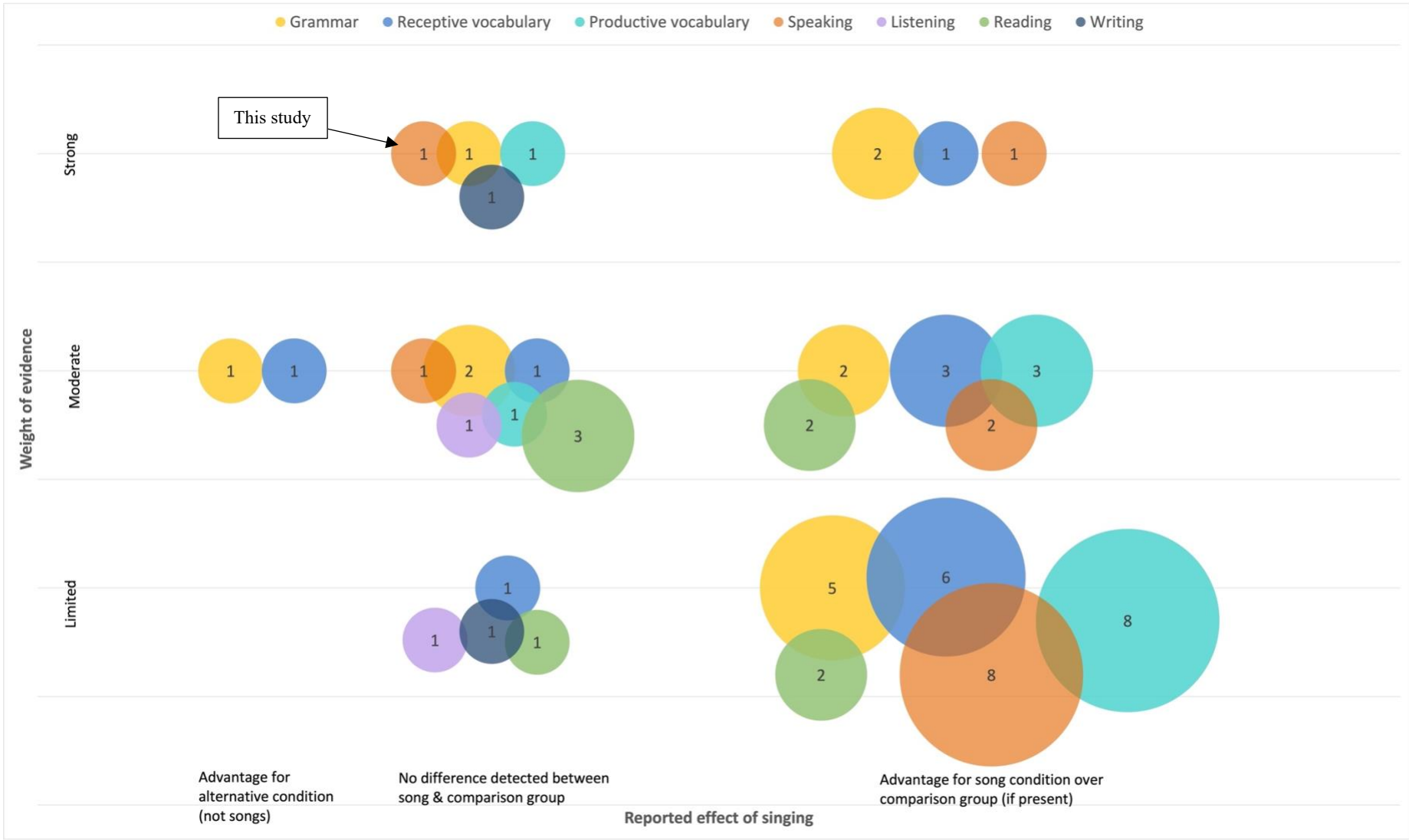
The systematic review in Phase 1 demonstrated considerable uncertainty about the effects of using songs compared to other teaching methods on YLLs' substantive linguistic outcomes. Those studies that found an advantage for singing in relation to a comparison group (either a control group or alternative treatment condition, where present) were more likely to be found at high risk of bias related to a variety of methodological limitations. In particular, studies often lacked clearly delineated research questions (making the precise aim of the study unclear, and therefore whether the study had addressed that aim), valid and reliable data collection methods, robust statistical analyses, and transparently reported outcomes. The

intervention in Phase 2 of this study only provides one small additional data point to add to our store of knowledge in the field. Figure 6.1 illustrates the position of Phase 2's intervention study in context with the included studies from Phase 1's systematic review, updating Figure 3.11 with the new study. It adds one more 'speaking' study (an orange dot) to the 'no difference detected' column, in the 'strong' weight of evidence row (since it is an RCT with a methodologically robust approach).

As is clear from the visualisation, overall, this study does little to redress the uncertainty uncovered by the systematic review. The majority of evidence is still of limited weight and overwhelmingly finds a positive effect of songs in comparison to alternative teaching methods or control conditions. Nonetheless, it shifts the evidence base slightly away from the positive end of the spectrum, particularly in the subset of speaking studies which are predominantly positive in their claims for songs. There are now two studies of speaking outcomes with low risk of bias: Chen (2011) found a positive relative effect of singing on the TOLD-P:3 measure of phonemic pronunciation, and my study found no evidence of difference between singing and comparison groups on an EIT measure. The jury is thus out on the overall effect of using songs to achieve L2 speaking outcomes. Songs appear to contribute to language learning, but not relatively more than alternative sources of linguistic input when speaking outcomes are measured.

The following section outlines implications for teaching practice based on the findings of Phase 2's intervention study.

Figure 6.1. The present study in context with the systematic review findings



6.6.1 Practice – use of songs for teaching FL is an evidence-based, personal choice

Phase 2's intervention study is (as far as I can ascertain) the first RCT conducted in England's primary school FL context to investigate the question of whether choosing songs (compared to alternative sources of language input and rehearsal) helps pupils achieve substantive linguistic outcomes such as "making substantial progress in one language" as mentioned in the national curriculum (DfE, 2013a). The data indicate that when new French input is presented through songs, and children are given the opportunity to listen to and repeat the songs orally, they can learn it and make progress. They do not, however, appear to learn more French or make more rapid progress through listening to and repeating songs than through listening to and repeating chants or stories, all else being equal. The key takeaway for teachers from Phase 2 of this thesis is therefore that a mixture of pedagogies is likely to be effective when presenting beginner French learners with new linguistic input. Whilst this is probably unlikely to be news to practitioners, who tend to express confidence in the idea that using multiple pedagogies for teaching languages is 'the best' approach (indeed, FL lessons are known for being an idiosyncratic mixture of activities – Finch et al., 2020), this study provides the first robust evidence that songs are approximately as effective for helping pupils make linguistic progress as chants or stories, which are likely to form part of the FL teaching repertoire too (DfE, 2013a; Hamilton & Murphy, 2023).

Secondly, in this study, the control group – who were presented with and rehearsed shorter, repetitive slot-filling structures – also improved their performance on the EIT. From this we could conclude that when learners are encouraged to notice the input through listening and produce output by speaking, they improve their ability to imitate the target language even for novel items that they have not encountered previously. Indeed, the control group had not encountered the EIT stimuli in their materials. The three experimental groups all performed similarly on RQ2c (the novel items subset) and not statistically differently to the control group, apart from the story group at delayed posttest. From these findings,

teachers could conclude that a focus on listening and repeating, with orthographic support through visuals (i.e., where written word forms are presented too; see Jiang, 2025) is likely to be an effective strategy for new learners of French to begin perceiving and producing the language, and that these skills will transfer to new linguistic contexts to a certain extent as well.

Whilst the data gathered here do not support any claims that songs constitute *more* effective material for FL teaching than other potential approaches such as reciting chants or stories, there is a clear case for stating that songs cannot be dismissed as merely a fun activity that are part of a repertoire of class activities but that do not directly support language learning outcomes themselves (e.g., activities for classroom behaviour management or affective purposes, see Hamilton & Murphy, 2023). As a teacher I believed the practice of using songs to teach FL was self-evidently useful and appeared to me, my colleagues, and on investigation a wider audience of UK-based primary teachers (Hamilton & Murphy, 2023), to be an effective approach to take with YLLs. However, given the strength of belief in the practice and anecdotal accounts of songs' effectiveness for achieving linguistic outcomes, it was alarming that almost no robust empirical evidence could be drawn upon to support the choice of songs over other approaches when the aim of an activity is to achieve a specific linguistic outcome, as opposed to an affective outcome such as motivating learners. This seemed counterintuitive. How could so many practitioners arrive at the same conclusion that songs are directly effective for achieving linguistic outcomes if so little of the research reliably found such an effect? This situation could have indicated that songs were not actually an effective pedagogy for achieving linguistic outcomes, and that other activities would make better choices for teachers introducing YLLs to the target language. Thus, one contribution of this study is to indicate that, whilst songs cannot be hailed as superlatively effective compared to alternatives, nor can they be dismissed as peripheral 'fun' activities that are not part of the 'real' teaching and learning.

This is good news for teachers who enjoy using songs. They can now choose to do so knowing that oral FL learning is certainly possible with such an approach, at approximately the same rate and magnitude as if they had chosen a story or chant (or by extension, poem or nursery rhyme) for their YLLs to listen to and rehearse. Equally, teachers who are less keen on singing need not fear that they are disadvantaging their YLLs by choosing alternative approaches: there is no evidence of a superlative effect of using songs, and it is therefore not a hindrance to learning to use fewer of them, if teachers prefer other approaches. If a teacher does not enjoy singing, they need not feel compelled to sing through fear of their learners missing out on the most effective pedagogy (which they would be forgiven for assuming is correct, given the strength of belief circulating in publications and materials aimed at teachers, see section 2.2). Again, a mixture of pedagogies is likely to be enjoyable for both teachers and pupils, and effective for helping pupils achieve substantive progress in the target language.

Furthermore, given the perceived precarity of music's place in the primary curriculum (Fautley et al., 2018), teachers could now choose to introduce more musical elements into the FL time available to achieve cross-curricular aims. Indeed, many primary teachers already do so (Fautley et al., 2018; Tinsley & Doležal, 2018). This study provides evidence that FL learning goals are indeed being met by making such a choice to join together music and FL activities. Further research could address the question of whether specific music curriculum learning outcomes are also being met by this kind of cross-curricular approach.

In summary, then, this study provides evidence that songs are one effective choice that teachers could make for helping beginner French learners achieve their first steps with learning the language orally, but it does not provide evidence that songs are a more effective choice than stories or chants for the introduction and rehearsal of new language. Teachers may thus choose to use songs as part of their teaching repertoire with confidence that they

are helping learners to make substantive progress in the FL that is comparable to the progress they would make if listening to and rehearsing stories or chants. This finding supports not only teachers' age-old intuitions about using songs to teach FL but also the many resources that include songs as part of the suggested materials to use with primary FL learners (DfE, 2011; Kirsch, 2008; see section 2.2) and to help them achieve "substantial progress" (DfE, 2013a) howsoever that is defined. It does not, however, answer the further question of how to take the knowledge of the song contents and bring it into productive use in a communicative context, as Conti (2015) and Cameron (2001) highlight. Further research will need to investigate this question in more detail.

Also, the important question of whether songs are more motivating for YLLs than other approaches, as often expressed by practitioners (e.g., DfE, 2013b), has not been addressed in this study. Now that we have a small additional amount of evidence that songs appear to address the basic assumption that learning is taking place when YLLs listen to and sing along with FL songs, future research could investigate the mechanisms through which FL could be learned in the most motivating way, which type of songs help learners make the most substantive progress, whether songs composed specifically for FL teaching are as facilitative of learning as traditional songs, and what the longer-term effects of learning FL through songs are compared to other approaches. The assumption that "something worthwhile is going on" (Rixon, 1991) when children sing FL songs is now slightly more well-substantiated by evidence, but the question of how much or why has yet to be answered.

6.6.2 Research – more and better research is needed

6.6.2.1 Volume of relevant research

Whilst Phase 1's systematic review found more numerous includable studies than previous reviews (Davis, 2017; Degrave, 2019; Engh, 2013; Sposet, 2008; Werner, 2020), 60 studies

over four decades does not amount to a large body of evidence. This is especially true given the heterogeneity of the research content: very few studies were found that investigate comparable outcomes for comparable populations using comparable methods. In terms of the extent of 'what works' evidence regarding using songs in FL teaching in primary schools, specifically in UK contexts, there is not so much a single gap as a wide-open field of possible future directions to take. For example, 35 studies in the systematic review investigated vocabulary outcomes. Ten of these were conducted in international primary school settings but none in the UK. There were ten studies investigating speaking outcomes conducted in primary school settings, one of which took place in the UK (Jarvis, 2013) but it is unclear whether this was in an EY reception class or a separate pre-school setting. The two other studies from UK settings were both secondary school studies looking at vocabulary (Legg, 2009) or vocabulary and grammar (Ludke, 2010) outcomes.

Songs are named in the English national curriculum as part of the exploration of "patterns and sounds of language" that contribute to pupils making "substantial progress in one language" (DfE, 2013a). Pre-dating the 2014 national curriculum, songs form a key element of the KS2 Framework (DfES, 2005a) recommendations for teaching approaches for achieving MFL oracy and literacy outcomes in primary schools. Yet, there is a vanishingly small amount of evidence (now including this project's own intervention study) that speaks to using songs in primary FL lessons in England's primary schools for achieving substantive linguistic outcomes. The potential of songs for helping YLLs learn languages has long been recognised by teachers. It is time for the research field to take up the challenge of testing what works when it comes to using songs to provide teachers with broader and deeper knowledge of their pedagogical choices.

Almost any starting point would be appropriate and potentially illuminating for teachers, given the scarcity of research in this area generally. Perhaps building on the work in this project would continue to deepen our understanding of how beginner learners in

England's primary schools could be introduced to new languages through rich, culturally relevant resources such as traditional songs, poems, chants and stories. There is much to be explored here in terms of the mechanisms through which more complex input (such as found in traditional song lyrics) compares to a more repetitive focus on form(s) approach, with slot-filling exercises that are incorporated into songs, as suggested in Kirsch (2008) for example. Many of the suggestions made by authors of teaching materials and approaches (see section 2.2.3) would bear further investigation through a 'what works' comparative research approach, as used in Phase 2's intervention. Whilst it may be challenging to conduct a RCT in school settings, the schools I worked with during Phase 2 were keen to be part of research that would inform them about the choices they make, rather than providing further anecdotal or contextual evidence. The inconvenience of randomisation was offset by the additional information they would gather to inform their practice decisions. Teachers already believe that songs 'work' for numerous linguistic outcomes; what remains to be seen is whether they work in different contexts, with different learners, and for different outcomes, and the mechanisms through which they work. Research of a larger scale and longer-term duration would build on the current evidence base, providing more information about the trajectories of learners across the KS2 years and comparing long-term outcomes for different approaches (with and without songs).

6.6.2.2 Methodological quality of relevant research

It is not only the volume of intervention research that is lacking: the quality of the research in this field needs to be greatly improved if teachers and policy makers are to draw upon robust evidence for making local and national decisions about primary FL pedagogy.

Only three of the 60 studies included in Phase 1's systematic review were assessed as bringing strong evidence (i.e., low risk of bias) to the debate. Fourteen studies were judged to have a moderate risk of bias and 43 studies were judged to have a limited global weight of evidence. This is a huge limitation in terms of drawing overall conclusions from included

studies, since a high risk of bias rating calls into question the trustworthiness of the conclusions in these studies. All efforts need to be made to minimise the risk of bias when conducting intervention research, particularly when claims are made about 'what works' in education. The first principle of such an approach is to make a fair comparison between groups, which may be achieved by randomly allocating participants to conditions so that potential effects of biases are reduced (Nunan et al., 2018). There is useful critique of claiming that randomised controlled trials in education (e.g., Deaton & Cartwright, 2018) are a supposedly 'guaranteed' way of generalising study findings beyond the study population, even when RCT quality measures are not met. Nonetheless, provided RCT assumptions are upheld (such as genuinely randomising participants to groups, not using an alphabetical or other pseudo-randomisation method), RCT designs permit researchers to all but rule out baseline biases as being responsible for any differences at outcome, and thus make more confident causal inferences associated with the interventions being compared (Campbell & Stanley, 1963; Connolly et al., 2017; Cook & Campbell, 1979; Gorard, 2003, 2013; Shadish et al., 2002; Slavin, 1986).

If, as a field, we are to move forward from Jolly's (1975) position that songs' potential benefits bear further study since "our intuitive feelings will remain only ideas unless they are, in some way, proven by means of study and experimental research" (Jolly, 1975: 14), then researchers must make every attempt to produce high quality, robust findings that can inform teaching practice. We should not shy away from trying to produce the highest possible standard of research to inform teachers of the evidence base for the practice choices they face. Particularly in this field, which is prone to folk theories and assumptions being recycled with ever-increasing certainty despite a lack of substance to claims (Bruner, 1996; Hamilton & Murphy, 2023), the position that Gleitman (1990: 3) outlines seems apt:

> One trouble with questions whose answers are self-evident is that investigators rarely collect the evidence to see if they pan out in practice. […] What is correct about such a position is by no means obvious, and therefore deserves serious study rather than acceptance as a background fact in our field.

In urging that 'serious study' be undertaken, I do not mean that all research in this field should consist of nothing more than randomised trials. Naturally, different research questions require different methods. There is no single set of standards for all research designs and, whilst I have taken a rigorous approach to what can be considered the 'gold' standard for experimental designs, all research designs play an important role in the wider research cycle.

Lieberman (2020) describes such a research cycle in the social sciences, outlining a process whereby a substantive research area is explored through early-stage observational studies of potential effects, correlational studies to establish association, small and larger-scale experimental designs to investigate causality, and replication studies to verify effects in additional contexts. This cycle is not necessarily linear as described here but accumulates in peer-reviewed publications over time. We might thus add systematic reviews to the cycle, to ascertain the state of the knowledge before embarking on new research to ensure it is necessary (Isaacs & Chalmers, 2023). Whilst all research forms part of the broad cycle of exploring a substantive topic, to establish the effects of one teaching method relative to others, experimental research requires (1) methodology that reduces uncertainty about the cause(s) of any putative effects and (2) transparent reporting of methods and data. Only once the full research cycle is adequately and reliably reported can we draw conclusions about what role songs may play in YLLs' FL development, and to which contexts and demographics the accumulated evidence is pertinent. Identifying where studies lie in Lieberman's (2020) research cycle could provide a useful framework for the field to begin locating knowledge gaps and gathering momentum by building on prior work. In this way, randomised trials form a part of the research cycle, and are complementary to observational

research, building on it to investigate potentials effects, and not a replacement for it. Indeed, the observations of centuries of languages teachers that songs appear to have a positive effect on YLLs' linguistic outcomes deserve a broad, coherent and robust programme of research to establish what the effects of songs are for different populations of learners in various contexts.

6.6.2.3 Reporting quality of research

A vital part of establishing a refreshed and stable evidential basis for teaching practice concerning the use of songs with YLLs is that research studies be reported clearly, transparently, and in a way that permits future replications to be conducted in novel contexts with additional populations of learners (Porte & McManus, 2019). Clear reporting also permits the end users of research, be they policy makers, practising teachers, materials designers, or researchers, to understand what research is being conducted and what the implications are for policy and practice. Even if a study were robustly conducted and met the highest standards of methodological rigour, without high standards of reporting nobody would know how trustworthy the methods and findings are.

In Phase 1's systematic review, poor reporting quality was responsible for a number of studies receiving moderate or high risk of bias assessments, simply because it is impossible to make informed judgements of the quality of research practice without clear reporting of the necessary information to form a judgement (Gorard, 2014). Whilst it is important to distinguish between quality of study design and quality of reporting (Moher et al., 1995), it is challenging to ascertain the quality of study design in cases where the reporting is compromised by incompleteness or loose expression that leads to vagueness about the details of a study. In the systematic review, for example, the method by which individual participants were allocated to conditions was described as 'random' in twelve studies (20% of included works) but seven of these did not report their allocation strategy

and four used diverse strategies that were not truly random (e.g., alternation based on alphabetisation of class lists).

Part of Deaton and Cartwright's (2018) concerns about randomised trials being used to make claims in educational research is the lack of true randomisation of the allocation process, as well as vague reporting standards. If we are to argue that RCTs represent a robust and reliable way of informing teachers of what works, as part of a wider research cycle (Lieberman, 2020), then as a field we must work towards designing, conducting and reporting research of a quality that is cumulatively valuable to practitioners. We will achieve this through our adherence to the highest standards, so that teachers can put their faith in research findings as they do the folk pedagogies that are verified through centuries of unwavering adherence to a common narrative of effective practice (Bruner, 1996). Journal editors also have their role to play in ensuring the highest quality of reporting standards are met by authors. They might do this by, for example, instructing authors to report RCTs according to CONSORT guidelines (Schulz et al., 2010) and systematic reviews according to PRISMA (Moher et al., 2009) to create a more transparent and unified approach to reporting.

6.6.2.4 Open science practices

Beyond establishing clear and high-quality reporting practices following genre-specific guidelines, the applied linguistics research community is moving towards more 'open science' culture and practices (Gass, 2019; Liu, 2023; Plonsky, 2024b). Open science relates to all stages of the cycle of research from pre-registering studies to open access publication. More than just the processes of research, open science is a philosophical commitment of the whole research system ('top-down') and an ethical consideration for researchers ('bottom-up'). Since we produce research with public stakeholders, we have a moral duty to share the outputs of that research openly with the public rather than behind paywalls. In terms of the processes through which research can be made more rigorous, transparent and replicable

which, as I have argued in this thesis, is necessary in this field, open science provides several opportunities at different points of the research process. Some key aspects are outlined here but see Plonsky's (2024a) edited volume for much more detail about open science in applied linguistics.

Pre-registration of designs, hypotheses, data analysis plans and code creates an *a priori* record of the planned processes that will be used during the research project. This can limit deviations from the plan whereby methodological choices are (perhaps accidentally or unconsciously) made to chase statistically significant findings where $p < .05$ (Forstmeier, Wagenmakers & Parker, 2017), or report the results of exploratory analyses as if there were confirmatory of *a priori* hypotheses (Nosek, Spies & Motyl, 2012), or other questionable research practices (Isbell et al., 2022). Such questionable practices – perhaps driven by publication bias towards 'significant' quantitative findings (Franco, Malhotra & Simonovits, 2014) – have arguably led to the so-called 'replication crisis' in the sciences broadly (Ioannidis, 2005) and also specifically in applied linguistics (Porte & McManus, 2019). Pre-registering a study (for example on the Open Science Framework; OSF, n.d.), or going further and adopting a registered report approach where papers are peer reviewed before data is collected and accepted for publication on that basis, regardless of the findings (Centre for Open Science, n.d.) provides a powerful way to create a paper trail for both accountability purposes and to facilitate replication and follow-up studies, thus expanding our knowledge of substantive topics and driving the field forwards because findings can be attributed to new populations rather than artefacts of methodology or analyses (Isbell, 2024).

Making data available on the OSF or the IRIS Database (Marsden et al., 2015) requires careful engagement with ethical good practice to avoid participants being identified, but it is a powerful tool for creating a transparent research process. When researchers make their data and analysis code available, analyses can be independently verified, studies can be replicated, and a deeper understanding of the research grows into new theories and

hypotheses to be tested (Gass, 2019; Plonsky, 2024a). Furthermore, making experimental materials openly available avoids "unnecessary duplication" (Isaacs & Chalmers, 2023:7), and moves research instruments towards greater validity and reliability (Isbell, 2024). As Isbell (2024) writes, "Simply put, when others can peer under the hood of a published research study, it is easier to understand exactly what was done and if it was done well, ultimately making it easier to trust the study's findings."

The principles of open science are aligned with teaching as an evidence-based practice. It is only through open and transparent conversations and collaborative efforts to work with teachers and other stakeholders that teachers can understand what research is aiming to achieve, and to what extent findings can a) be trusted and b) are pertinent to their particular learners and context. If we are to avoid the gloomy state of teaching becoming "merely the transmission of self-perpetuating, unsupported beliefs and prejudices" (Paran, 2017: 506), we must bring teachers' experience, expertise and intuition together with the best available external research findings (Chalmers, 2016). Open science principles give research a better chance of being accessible to and accepted by teachers who might then see the value of going directly to read research rather than relying on secondary or possibly biased accounts of the effects of using songs for achieving FL outcomes from materials developers or social media sites.

6.6.3 Theory

The findings of the systematic review were inconclusive regarding the causal relationships between using songs with YLLs and their linguistic learning outcomes. It was not possible to claim, based on the collected studies, that there either is or is not a reliable effect of introducing FL through songs and language acquisition of learners in formal education settings. Furthermore, as discussed in section 2.3.1, some of the theoretical motivation for assuming there is a link between songs and linguistic outcomes in formal educational settings rests on unstable evidential foundations. Whilst teachers (Hamilton & Murphy,

2023; Murphey, 1990) and researchers (Engh, 2013; Fonseca-Mora, 2000; Paquette & Rieg, 2008) draw upon theories that themselves require more substantiation, there is little hope of moving the field forward without better operationalisation of theories to underpin experimental research. The studies by Campfield and Murphy (2013; 2014) drew upon the L1 theory of prosodic bootstrapping, a well-evidenced hypothesis with some explanatory power for the way infants and young children parse speech streams and begin bootstrapping the lexis and syntax of their first language(s). Phase 2's intervention did not test prosodic bootstrapping and the relative prosodic characteristics of songs, chants and stories directly. Yet the findings from RQ2c are consistent with the theory that children can exploit the phonological form of novel L2 input, using their existing or emerging mental representations of L2 syntax and lexis to parse, process and produce the EIT stimuli containing novel items of lexis. Future work could take these starting points further by testing different L1/L2 language combinations, and different age groups of learners, or using methods from psycholinguistics such as eye tracking to ascertain the extent to which participants rely on the visual support of the orthographic input to parse the speech stream compared to just listening to the input orally, for example.

This research, however, is more focused on the pedagogical value of using songs with YLLs in FL primary school lessons. Given the strength of belief in songs' effectiveness for teaching YLLs a new L2, it is encouraging to know that the first and most basic link between singing songs and improved linguistic performance has been tentatively demonstrated. There was no indication in the data of songs leading to superlative performance in the EIT outcome measure over time. Future studies might ask whether, beyond the six-week delayed posttest, there is any relative difference in performance between experimental groups to test the hypothesis that songs are more memorable than other forms of FL classroom input. It could be that this study did not detect such an effect due to the delayed posttest occurring relatively soon after the end of the three-week

intervention. A delay of several months might find different effects and then examine the mechanisms through which songs get 'stuck in our heads' (Murphey, 1990).

More importantly, perhaps, this study has provided some evidence for teachers who enjoy using songs with their YLLs to indicate that this valued practice does what teachers think it does in terms of learners improving their L2 performance on an oral measure of proficiency such as the EIT (Davanellos, 1999; Forster, 2006; Garton et al., 2011; Hamilton & Murphy, 2023; Harris & O'Leary, 2009; Linse, 2006; Paquette & Rieg, 2008; Saricoban & Metin, 2000; Schoepp, 2001; Şevik, 2011; Walker, 2006). The next steps are to assess how and why songs might have such an effect on L2 outcomes, and to build a clearer picture of what purposes teachers might choose to use songs for. In the KS2 Framework (DfES, 2005a), songs are clearly linked with oracy and literacy outcomes. There is a vanishingly small amount of evidence on this topic in L2 (as demonstrated throughout this thesis) and also in L1 contexts (e.g., Lonie, 2010, found very little reliable evidence to link songs to literacy outcomes in a review of the research). To build up a clear picture for theory, then, a clear and cumulative programme of research is required that slowly pieces together the puzzle of what songs contribute to L2 learning outcomes, how they contribute (what forms of songs and in what doses or circumstances?), and why. An interdisciplinary approach to the topic might be valuable, given the broad nature of such questions which could also include the intercultural value songs bring to FL learning and affective outcomes.

Thus, whilst this thesis provides only one small piece of the puzzle, it is a corner piece from which we might slowly but surely begin to collect and build the rest of the picture. Teachers deserve the highest standard of meticulous research across the entire research cycle (Lieberman, 2020) to be carried out to help them make the best and most theory-informed choices for their YLLs, in tandem with their experiential and practice-informed knowledge (Chalmers, 2016). An important part of the theory-building process is for research to build on prior work, as Phase 2 attempts to build on Campfield and Murphy

(2014) and Davis and Fan (2016). In doing so, we can begin asking more nuanced and precise research questions, and begin narrowing down what is currently a wide-open question about the nature of the relationship between FL songs and linguistic outcomes.

In Phase 1's systematic review, there was little evidence of included studies building on the foundations laid by previous similar studies. The research questions (albeit only where present, since 23 of the 60 studies did not report clearly-defined research questions) tended to repeat similar 'starter' questions about investigating the effectiveness of using songs, with little reference to prior research, nor clear articulations of the theoretical basis for asking those questions. Language outcomes were sometimes named specifically in research questions but were often left open. For example, the question "Will a music program designed to be incorporated into the second language program of Grade 2 French immersion students enhance the learning of both music and language?" (Lowe, 1995) does not narrow down the focus of study to any substantive aspect of language. And whilst a question such as "Is teaching grammar using songs effective to improve students' grammar?" (Alinte, 2013) specifies 'grammar' is the outcome under study, there is no indication of which aspect of grammar is being investigated. More precision would help move the field forward.

As well as outcomes, interventions need to be clearly specified. Whilst some research questions from included studies name a highly specific intervention (e.g., "Are finger family collection YouTube videos nursery rhymes impact on Iraqi EFL Pupils' Performance in speaking skills?" Al-Mosawi, 2018), others do not narrow down the authors' definition of song at all (e.g., "Does the use of song have any effects on vocabulary retention of preschool young English language learners?" Madani & Nasrabadi, 2016). The possibilities for using songs in FL lessons is broad, as noted in section 1.1. It is therefore essential for researchers to clarify exactly how they used songs in their interventions, so that theories can specify the proposed nature of the relationship between using songs in FL lessons and clearly defined

linguistic outcomes. Then teachers would be able to draw upon the research evidence base to know for what purposes to use songs and how to use them.

Finally, some included studies report research questions that 'drift' from one version to another within papers. In her master's thesis, Siebring (2004: 12) initially asks, "Does the use of a systematic approach based on songs help prevent or correct errors made by students of FSL?" which becomes, "Can songs that target specific errors be used to prevent or correct those errors in Core French?" (Siebring, 2004: 95) later in the thesis. Research questions' shifts in focus, lack of specific focus, or narrow and apparently singular focus is possibly symptomatic of the field lacking coherence and direction. Or, simply, it could indicate a need for clear expectations and training for researchers on how to write and report research questions. Since many peer-reviewed included studies also lacked clearly defined research questions, however, the issue cannot be solely attributed to students' contributions to the field. Either way, the challenge moving forward will be to pull the field up by its bootstraps and create a firm foundation on which to build future research that can contribute to our collective theoretical understanding as well as to teachers' pedagogical practice in a more robust way.

6.7 Limitations of the study

The intervention in Phase 2 of this study has a number of limitations that should be considered when interpreting the findings of the study, as well as addressed in any replication or extension studies.

6.7.1 Multimodal input

With the aim of following as closely as possible the methods in Campfield and Murphy (2014) to build on their work, this study also presented learners with audio and visual (multimodal) content when introducing them to new FL input during the three-week intervention (see section 4.4.3.4, on presentation of input materials). As well as listening to and repeating the input, participants could see illustrations accompanying the words they

were rehearsing. There is evidence that presenting written word forms alongside aural input visually supports learners' oral language development (Jiang, 2025). The EIT outcome measure was an oral measure of language proficiency where participants were asked to listen to and repeat the stimuli without any visual support. Thus, there was an additional modality in the input, the written form, that did not appear in the outcome measure, otherwise the input and outcome modalities were the same (listening and speaking). Since Phase 2's study was designed to expand Campfield and Murphy (2014), it was necessary to use the same method of presentation to make a useful comparison of the two studies.

Arguably, if prosody supports young children in parsing the speech stream (see section 2.3.1.3 on prosodic bootstrapping), then it is possible that presenting the written word forms in this cross-modal manner reduced the potential effect of the two prosodically salient conditions relative to the story condition by making word boundaries visibly salient in all conditions. If no written word forms had been presented, the song and chant conditions may have produced a larger effect relative to the story condition since, in the absence of visual support, the more salient prosody of the songs and chants may have been more facilitative of parsing the speech stream. Future research could test the effect of providing additional visual support on primary FL learners' relative performance across these three input conditions.

However, it seems unlikely to be ecologically valid to exclude visual written support at the primary school level for the extended period of the intervention, due to the importance placed on literacy in the primary curriculum (DfE, 2013a). With younger pre-literate children in early years settings, the written forms would perhaps not be presented to accompany songs or chants. The early years context might therefore be an ecologically valid way of comparing songs, chants and stories through listening and speaking modes. In that case, though, any findings gathered from pre-literate early years participants would not automatically apply to primary school FL learners who are literate.

6.7.2 Statistical power

As noted in section 4.3.2.3, this study was powered to detect a small effect (.15) with *alpha* = .05, and *beta* = .8 for a factorial 4*3 ANOVA, which does not fully account for the complexity of mixed-effects modelling that was undertaken. A revised power calculation suggests the study sample size (n = 96) was insufficient to detect effects smaller than .39. Therefore, if the intervention had a small but meaningful effect, the study was unlikely to have enough power to detect it and has an increased likelihood of Type II error (failure to detect an effect that is present).

Cases where differences in outcomes between groups were not statistically significant, as in the CLMM analyses in this study, cannot be interpreted as evidence of no effect, just that the study lacked sufficient power to detect one if one was present. Simulation-based power calculations would provide a more precise approach for mixed-effects modelling power calculations, but it is challenging to estimate realistic expected effect sizes since there is such a paucity of robust research in the field (as discovered in Phase 1's systematic review) to base estimations of effect size on.

If the goal of future research is to power for the detection of very small interaction effects, researchers should aim to recruit a larger sample size to increase statistical power. Since, however, this study involved calculating Bayes factors, and these also returned no evidence of any difference between the song condition and chant or story conditions, it seems unlikely that the effect is of a magnitude that would produce meaningful differences in outcomes in a real-world classroom. Thus, in spite of being underpowered to detect very small interaction effects, this study nonetheless contributes meaningfully to our understanding of the relative effects of using songs, chants and stories in FL lessons on YLLs linguistic outcomes.

6.7.3 Selection bias

Ideally, to generate the most robust and generalisable findings (Gorard, 2001), random sampling from the population of Y3 pupils in state-maintained primary schools would have formed the sample for this study. Since true random sampling was not possible, I used a mixture of purposive and convenience sampling and invited ten large local primary schools to take part to achieve as representative a sample of the population as possible. Two invited schools volunteered to participate. There is thus selection bias at the level of the school (Catalogue of Bias Collaboration, 2017) since they may differ systematically from those schools that did not volunteer. I cautiously propose that the two participating schools (one suburban, one inner city) make them broadly representative of FL learners in state-maintained primaries by representing different socio-economic and demographic contexts. Both schools, however, have two very keen linguists leading their Y3 FL programmes, which may set them apart from primary schools without similar human resources. Future studies could attempt to randomly sample at the school level, but this is a challenge no matter how large or well-funded a study is due to the difficulty of compiling a complete list of all Y3 pupils in the country.

At the participant level, parental opt-in consent was required for children to take part in the study. It is therefore possible that systematic differences exist between parents (and by extension their children) who completed the forms and allowed their children to participate, and those who did not. Both schools endeavoured to encourage as many eligible students as possible to participate by gathering parental consent through their online communication systems, providing paper copies of consent forms, asking children to remind their parents to return forms, and speaking to parents at the end of the school day. School 1 had fewer total participants than School 2, and a larger proportion of School 2 parents returned signed consent forms. These differences may indicate biases in the selection process.

Perhaps parents who opted in have a particular interest in their children studying languages or taking advantage of additional activities, and those who did not opt in are less keen on their children studying languages or taking part in additional activities (even though the intervention was a replacement for their planned FL lessons). The participants may have differed in their levels of support received at home for language learning and been more enthusiastic than children who did not participate, limiting the applicability of the findings to a more diverse population of learners. Participants were asked for their oral assent to take part in the screening variables and EIT measures. A participant in School 1 refused to take part in the EIT audio recording at any of the three time points, and was subsequently removed from data analysis. This could have biased the data slightly towards children who voluntarily took part, who may have a greater enthusiasm for or ability to participate in the tasks.

One way of reducing selection bias at the participant level would be to make it possible for school head teachers who read the study information sheet and agree for their school to participate in research studies to be the 'gatekeeper' for all children in their school. Chalmers (2019) argues that this would be preferable to the *un*controlled experiments that take place every day in classrooms all over the world, whereby teachers are at liberty to adopt new and untested teaching approaches without the need for parental or student consent, and without formally evaluating their effects relative to previous practice or the variety of available alternatives. It is surely preferable to have a new approach formally and carefully evaluated against a comparison in a fair test of their relative effects than to adopt a new approach without fully evaluating its effects in fear, for example, that some children will 'miss out' on a potentially beneficial approach. Since there is no way of knowing whether a new approach is beneficial without a fair test, this seems like a cyclical and misguided argument. Trials would have increased statistical power and be less prone to selection bias if headteachers' consent could be considered sufficient for their pupils to take

part in a study, much as the everyday business of schools' approaches to teaching and learning is decided upon by headteachers and leadership teams. If education research were better powered to detect effects when they are present, this would be beneficial in helping teachers make decisions about the most effective pedagogy to choose for their learners and intended learning outcomes.

6.7.4 Cross contamination

Participants from the four study conditions were from two schools and randomly allocated from different classes into the intervention conditions. Outside of the intervention periods, therefore, they interacted with children participating in the comparison conditions. This poses a threat to the study's internal validity since participants could have discussed the input they were receiving or even taught each other some of the materials from their condition. There may also have been unequal cross-exposure. The narrative arc of the story could be shared more easily than the exact words, but a catchy song is easily shared verbatim, potentially leading to phrases from the songs being shared more than phrases from other conditions. Additional learning through repetition outside the intervention classes could have reinforced some children's understanding of the EIT stimuli more than others, potentially confounding the results. Any of these possible routes into cross contamination would make it difficult to determine whether potential effects between conditions were diluted, which could explain why there was very little difference in performance on the EIT detected between conditions.

Whilst it would not be realistic to isolate groups from each other entirely for the three-week period of the intervention, I attempted to reduce possible cross contamination by explaining to the participants about the 'fair comparison' we were trying to make in our science project together. I explained that, even if they really wanted to share what they were doing with their friends, it would be good if they could wait until after we had finished doing our research together and then enjoy sharing all the materials in their classes. This hopefully

discouraged explicit sharing, without making it too restrictive and unnatural for the children. There were a couple of very enthusiastic participants, but they were as enthusiastic about the importance of doing real research as they were about taking part in French lessons, so they seemed to understand why it was important to keep each group's activities separate. More practically, the intervention lessons in both schools took place in rooms that were in a different part of the building from the classrooms. It was therefore not possible to overhear the intervention lessons from the classrooms. Therefore, if some cross contamination occurred by informal sharing of materials, this is unlikely to have entailed a full and structured form of exposure equivalent to the input in each condition. The majority of each participant's exposure would still derive from their own condition and that would be the dominant influence on their performance.

6.8 Summary

There is a widespread belief that songs represent a powerful and even superlatively effective pedagogy for teaching children second or foreign languages. This assumption is based upon centuries of teacher observation and anecdotal evidence. Phase 1 of this project found scant existing research to reliably substantiate any causal claims about songs' effects on YLLs' linguistic outcomes. Phase 2's randomised trial of the relative effects of using songs to introduce new French input compared with the alternative approaches of using chants or stories found no statistically significant evidence to support claims that songs are more advantageous than other methods for achieving L2 progress with beginner French learners. It did, however, find that songs produce comparable effects to chants and stories. These findings indicate that teachers should choose the pedagogical approach they are most attuned with for introducing French to their beginner learners in Y3 FL lessons. None of the approaches studied was found to be more or less beneficial than the others, on average, in this study. Nor can it be claimed that songs, chants and stories are identical in their effects on

beginner primary school learners' acquisition of French: for such a statement to be made legitimately, more evidence is required.

Chapter 7

Conclusion

The practice of using songs to introduce new languages to young learners is centuries old (Murphy, 1968), and valued by teachers internationally (Garton et al., 2011; Harris & O'Leary, 2009; Linse, 2006; Şevik, 2011) and in the UK (Hamilton & Murphy, 2023). Teachers of YLLs use songs in multiple ways including as individual, small-group, or whole-class singing activities, and as listening activities presented via screen, audio recording, or live performance (Hamilton & Murphy, 2023); as gap-filling, sequencing, grammar or vocabulary exercises, or stimuli for creative output (Davanellos, 1990); and in all four modes of listening, reading, speaking and writing (Conti, 2015; Walker, 2006). Songs are used for their enjoyable, calming and motivating effects (Jolly, 1975; Rumley, 1999; Smith, n.d.), and for introducing cultural or authentic FL content (Jones & Coffey, 2016), especially when teachers lack confidence in their own language teaching skills (Graham et al., 2017; Unopeia, 2022).

Importantly, songs are perceived by policymakers and practitioners as being highly effective for achieving linguistic outcomes such as promoting pupils' L2 oracy and literacy skills (Cameron, 2001; DfE, 2013a; DfES, 2005a; Forder et al., 2013; Kirsch, 2008), making connections between languages (Mordsley, 2017), and achieving linguistic progress in vocabulary, grammar, and pronunciation (Hamilton & Murphy, 2023; Paquette & Rieg, 2008; Forster, 2006; Walker, 2006). Confident causal claims abound in teacher-facing resources about using songs in various guises to achieve these FL linguistic outcomes (e.g., Davanellos, 1999; Forder et al., 2013; Kirsch, 2008; Paquette & Rieg, 2016; Thain, 2010). In short, using songs in FL teaching is a ubiquitous and well-loved practice with plenty of intuition and experience testifying to the potential benefits of choosing such a pedagogical

approach with YLLs. As noted fifty years ago, however, "our intuitive feelings will remain only ideas unless they are, in some way, proven by means of study and experimental research" (Jolly, 1975: 14). As far as I could ascertain as a practising teacher and then master's student, such experimental research had not provided the evidence that I and other teachers sought to underpin our practice regarding using songs in FL lessons.

This situation is problematic because young learners merit the highest quality research to be carried out so that their teachers can base their pedagogical judgements of what will work best for particular outcomes in specific contexts on trustworthy and reliable evidence. Teachers are experts in accruing the kind of skills, expertise and cross-situational insights that give them a keenly developed intuition about which pedagogical choices to make to meet particular educational ends for particular learners. The kind of unquestioning faith in songs I uncovered in my master's study (Hamilton & Murphy, 2023) seemed to suggest songs were a panacea for all manner of FL outcomes and classroom purposes (e.g., Forster, 2006; Paquette & Rieg, 2008). No matter how evidence-based practitioners aim to be when making pedagogical choices, teachers can only base their professional judgements on evidence from research if such research exists, and if it is accessible (Borg, 2009). Perhaps it did exist and I had just failed to find it.

The purpose of this Doctoral study was therefore two-fold. Firstly, prompted by the mismatch between many FL teachers' faith in the benefits of using songs to teach YLLs but the apparent scarcity of research into the topic, it aimed to provide a robust and dispassionate appraisal of research investigating the second or foreign language-learning benefits of using songs to teach YLLs. This aim was achieved by conducting a systematic review and narrative synthesis to assess existing evidence about songs' effectiveness as language-learning tools in second or foreign language classroom contexts with YLLs. As discussed in the systematic review, there are limited numbers of well-controlled and rigorous studies that examine the substantive linguistic effects of presenting FL input in the form of songs

compared to alternative approaches. Studies that do so have found equivocal results across different participant demographics and linguistic outcomes. There is no way of knowing, from the assembled literature in Phase 1, what the causal effects of using songs on YLLs' linguistic outcomes are, nor the extent to which any effects of using songs may differ compared to other teaching approaches. Compounding this issue is the fact that, of the 60 studies included in the systematic review, only three received low risk of bias assessment ratings. These three studies did not investigate the same linguistic outcomes with the same populations of learners or introduce songs in the same form or manner (or introduce them at all, in the case of Campfield and Murphy, 2013, whose study included nursery rhymes but not songs). Even if they did have more comparable approaches to the questions they addressed, three studies is limited in terms of what can be concluded about the general state of learning for all YLLs in FL educational contexts. It was clear at the end of Phase 1 that much more, and much better, research needed to be conducted that would permit reliable and valid causal conclusions to be drawn about the relationship between songs and FL learning.

Then, based on the current state of our collective knowledge on the topic as assessed in Phase 1's systematic review, this project's second purpose was to design and conduct a rigorous intervention study to contribute novel evidence to the field. Phase 2's intervention study examined the relative effects of presenting and rehearsing French input through the form of songs, chants or stories on 7–8-year-old beginner learners' acquisition of French in two English primary schools. To my knowledge, it is the first randomised controlled trial to be conducted in the UK (and possibly the world) investigating the substantive linguistic outcomes of songs compared to other frequently-used teaching methods with primary school-aged children. The findings link the use of songs, chants and stories with the participants' improved performance in the EIT outcome, a measure of oral language competence (Erlam, 2006). This study is the first of its kind to isolate the effects of songs on primary FL learners' oral language outcomes, relative to other age-appropriate teaching

approaches. It builds on the work of Campfield and Murphy (2014) who found that Polish EFL primary learners improved their performance in an EIT outcome when learning nursery rhymes or stories. Taken together, these studies indicate that teachers might choose a mixture of pedagogies including songs, rhymes, poems, chants and stories to rehearse as whole-class oral language activities. The data do not indicate that songs are more effective than other ecologically valid alternative approaches over the three-week time frame in this study, despite cultural beliefs that songs are superlatively effective for learning languages because they are highly memorable and enjoyable. Further research will need to investigate the mechanisms through which songs might contribute to language-learning. Longitudinal studies could look at longer-term memories of input encountered across different conditions, or the extent to which songs increase YLLs' motivation levels for sustained language learning, especially as they make the challenging transition from primary to secondary school FL (Burstall et al., 1974; Cable et al., 2012; Hunt, 2009; Macaro, 2008; Wade et al., 2009). For now, teachers might choose songs if they wish to use them in the knowledge that they are part of a sound approach for introducing French (and arguably, by extension, other FLs) to primary school learners. Equally, if teachers are less interested in using songs, they need not feel that they are disadvantaging their learners by choosing other approaches such as chanting or reciting stories aloud more often than singing songs together.

Looking forward, the field would benefit from a coherent and strategic research agenda. Using songs with YLLs is clearly a practice that many primary FL teachers engage with and value. But there is a discrepancy between strength of belief in the practice, and strength of evidence to underpin it from empirical research. If multiple studies of a rigorous and similar nature were conducted, it would be feasible to conduct a systematic review and potentially a meta-analysis to gauge overall effect sizes (Petticrew & Roberts, 2006). The explanatory power of cumulative evidence, especially when that evidence is cumulated scientifically through reliable forms of research synthesis, collaboration and open science

practice such as data sharing (Chalmers et al., 2022) is not a nice addition to the research cycle where questions of causality are concerned. It is the foundation of good scientific practice. Our teachers and learners should demand the best possible empirical evidence for informing the teaching and learning process, and researchers should endeavour to provide them with it. This project has contributed to the ongoing research cycle by gathering evidence upon which teachers can base their pedagogical choices when considering using songs for FL teaching and by which new research can be informed. There remains much more work to be done to investigate this valued practice with greater empirical rigour and from different conceptual angles. Replication in different contexts and with new populations of learners (for example, in the early years or secondary schools) would be welcome, to further interrogate the evidence gathered here and to move the field forward one small step at a time (Porte & McManus, 2019). Perhaps then, researchers and teachers will start singing from the same songbook for the benefit of young language learners.

References

Asterisk (*) denotes study is included in the systematic review.

Abboub, N., Nazzi, T., & Gervain, J. (2016). Prosodic grouping at birth. *Brain and Language*, *162*, 46-59.

Abrams, R., & Gerhardt, K. (2000). The Acoustic Environment and Physiological Responses of the Fetus. *J Perinatol, 20*(Suppl 1), S31–S36. https://doi.org/10.1038/sj.jp.7200445

*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

Alviar, C., Sahoo, M., Edwards, L.A., Jones, W., Klin, A., & Lense, M. (2023). Infant-directed song potentiates infants' selective attention to adults' mouths over the first year of life. *Developmental Science, 26*(5), e13359.

Allen, G.D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In G.H. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.). *Child phonology. Volume 1: Production* (pp. 227– 256). New York, NY: Academic Press.

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

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

*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, Especial No.17(2018), 452–474.

*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.

ana_loves_this_world (2023, 23 May). Comment on knoxstudy (2023) [social media post] Newborn babies have accents! Knox Study OGs may remember this from 2020. #language #accents #interesting #todayilearned. *Instagram*. Accessed on 12th January 2025 at https://www.instagram.com/reel/CsmE29Jol1B/?igshid=MmJiY2I4NDBkZg%3D%3D

Anthony, L., & Nation, P. (2021) *Picture Vocabulary Size Test, Build 1,2,3*. Accessed 24th July 2024: https://laurenceanthony.net/software/pvst/releases/PVST123/help.pdf

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

Arvaniti, A., & Fletcher, J. (2020). The auto segmental-metrical theory of intonational phonology. In C. Gussenhoven & A. Chen (Eds.). *The Oxford Handbook of Language Prosody*. pp.78–95. Oxford: Oxford University Press.

Aslin, R.N., Woodward, J.Z., LaMendola, N.P., and Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K.Demuth (Eds.) *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, pp.117–134. Mahwah: Erlbaum.

*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.

Axehandle1234. (2022). *Thoughts on the teaching of MFL in primary*. [Online forum post]. Reddit. Accessed on 16th December 2024: https://www.reddit.com/r/TeachingUK/comments/xznhe4/thoughts_on_the_teaching_of_mfl_in_primary/?rdt=35891

Aznavour, C. (1965). La Bohème [Song]. On *Monsieur Carnaval*. Barclay Records.

Baguley, T., & Kaye, W. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by Z. Dienes. *British Journal of Mathematical and Statistical Psychology, 63*(3), 695–698.

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

Baker, C.L. (1979). Syntactic Theory and the Projection Problem. *Linguistic Inquiry, 10*(4), 533–581.

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.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

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

Ben-David, B.M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: test for rating of emotions in speech. *Journal of speech, language, and hearing research, 59*(1), 72–89.

Benavides-Varela, S. & Gervain, J. (2017). Learning word order at birth: A NIRS study. Developmental Cognitive Neuroscience, 25, 198–208. https://doi.org/10.1016/j.dcn.2017.03.003

Bergeson, T. R., & Trehub, S. E. (2002). Absolute pitch and tempo in mothers' songs to infants. *Psychological science, 13*(1), 72–75.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child development, 89*(6), 1996–2009.

Bernard, C., & Gervain, J. (2012). Prosodic Cues to Word Order: What Level of Representation? *Frontiers in Psychology*, 3, 451. https://doi.org/10.3389/fpsyg.2012.00451

Bishop, D.V.M., & Thompson, P. (2023). *Evaluating What Works: An Intuitive Guide to Intervention Research for Practitioners (1st ed).* NYC: Chapman and Hall/CRC. https://doi.org/10.1201/9781003453079

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. M. Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (pp. 41–68). Cambridge: Cambridge University Press.

*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.

Bolton, T.L. (1894). Rhythm. *The American Journal of Psychology, 6*(2), 145-238.

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

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition, 65*(1), 33–69.

Broesch, T.L., & Bryant, G.A. (2015). Prosody in infant-directed speech is similar across western and traditional cultures. *Journal of Cognition and Development, 16*(1), 31–43.

Brumfit, C., Moon, J. & Tongue, R. (1991). *Teaching English to Children: From Practice to Principle.* London: Collins.

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

Bryant, P. E., Bradley, L., Maclean, M., & Crossland, J. (1989). Nursery rhymes, phonological skills and reading. *Journal of Child Language*, *16*(2), 407–428. https://doi.org/10.1017/s0305000900010485

Bryant, G.A., Liénard, P., & Clark Barrett, H. (2012). Recognizing infant-directed speech across distant cultures: evidence from Africa. *Journal of Evolutionary Psychology, 10*(2), 47–59.

Burstall, C., Jamieson, M., Cohen, S. & Hargreaves, M. (1974). *Primary French in the Balance*. Slough: NFER.

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

Byers-Heinlein, K., Burns, T. C. & Werker, J. F. (2010). The Roots of Bilingualism in Newborns. *Psychological Science, 21*(3), 343–348. https://doi.org/10.1177/0956797609360758

Cable, C., Driscoll, P., Mitchell, R., Sing, S., Cremin, T., Earl, J., Eyres, I., Holmes, B., Martin, C., & Heins, B. (2012). Language learning at Key Stage 2: findings from a longitudinal study. *Education 3-13, 40*(4), 363–378. https://doi.org/10.1080/03004279.2012.691371

*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.

Cameron, L. (2001). *Teaching Languages to Young Learners*. Cambridge: Cambridge University Press.

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. (2010). *Factors affecting early second language acquisition: the role of linguistic rhythm.* Doctoral thesis, University of Oxford.

*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 acquisition. *System, 42*, 207–219. https://doi.org/10.1016/j.system.2013.12.002

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, I*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Catalogue of Bias Collaboration, Nunan, D., Bankhead, C., Aronson, J.K. (2017). Selection bias. *Catalogue Of Bias.* Available at: http://www.catalogofbias.org/biases/selection-bias/ [Accessed 24 Feb 2025].

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain online recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, 10, 1–18.

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.

Centre for Open Science (no date). *Registered Reports: Peer review before results are known to align scientific values and practices.* [Accessed online 26th February 2025]: https://www.cos.io/initiatives/registered-reports.

*Chae, Y., & Yoon, E. (2013). 채영신; 윤은자 (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. (2019). *Leveraging the L1: The role of EAL learners' first language in their acquisition of English vocabulary*. Doctoral thesis, Oxford Brookes University.

Chalmers, H., Brown, J. & Koryakina, A. (2024). 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*, *15*(4), 1645-1669. 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

Chalmers, I., Bracken, M.B., Djulbegovic, B., Garattini, S., Grant, J., Metin Gülmezoglu, A.M., Howells, D.W., Ioannidis, J.P.A., and Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet, 383*(9912), 156–165.

Chalmers, I., Hedges, L.V. & Cooper, H. (2002). A Brief History of Research Synthesis. *Evaluation & the Health Professions, 25*(1), 12–37. https://doi.org/10.1177/0163278702025001003

*Cheippe, E. (2012). *La voie musicale pour remédier aux di*ffi*cultés de prononciation des voyelles de l'allemand dans des textes lus: expérimentation dans une classe bilingue: analyse acoustique.* Doctoral thesis, Université de Strasbourg.

Chen, A., Esteve-Gibert, N., Prieto, P., & Redford, M.A. (2020). Development of phrase-level prosody from infancy to late childhood. In C. Gussenhoven & A. Chen (Eds.) *The Oxford Handbook of Language Prosody*, pp.553–562. Oxford: Oxford University Press.

*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*. Doctoral thesis, University of Florida.

Chen, J.L., Penhune, V.B., & Zatorre, R.J. (2008). Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical

training. *Journal of cognitive neuroscience*, *20*(2), 226–239. https://doi.org/10.1162/jocn.2008.20018

*Chiang, M. (2003). *The effect of chanting activities on the comprehension of English of first graders and college freshmen in Taipei*. Doctoral thesis, Texas A&M University-Kingsville.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

Chomsky, N. (1980). *Rules and Representations.* New York: Columbia University 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

Christensen, R. (2023). *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1, https://CRAN.R-project.org/package=ordinal

Christophe, A., Guasti, T., Nespor, M., Dupoux, E., & Van Ooyen, B. (1997). Reflections on Phonological Bootstrapping: Its Role for Lexical and Syntactic Acquisition. *Language and Cognitive Processes, 12*(5–6), 585–612. https://doi.org/10.1080/016909697386637

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 December 6, 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.

Collen, I., & Duff, J. (2024). *Language Trends England 2024.* UK: British Council.

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. Los Angeles: SAGE.

CONSORT (no date). Welcome to the CONSORT Website. Available at: http://www.consortstatement.org Accessed 10 June 2024.

Conti, G. (2015). *How to exploit the full learning potential of an L2 song in the language classroom*. [Blog, 15th June]. Accessed on 14th December 2024: https://gianfrancoconti.com/2015/06/15/how-to-exploit-the-full-learning-potential-of-a-target-language-song-in-the-mfl-classroom/

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Chicago: Rand McNally College.

Cooper, R.P., & Aslin, R.N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*(5), 1584–1595.

*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*. Doctoral thesis, University of North Carolina.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

*Cruz-Cruz, M.L. (2005). *The effects of selected music and songs on teaching grammar and vocabulary to second grade English language learners*. Doctoral thesis, Texas A&M University, Kingsville.

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.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, 92, 81–104. https://doi.org/10.1016/0024-3841(94)90338-7

Cutler, A., & Foss, D.J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, 21, 1–10

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

de Carvalho, A. (2017). *The role of phrasal prosody and function words in the acquisition of word meanings*. Doctoral thesis, Université Paris sciences et lettres.

de Carvalho, A., He, A. X., Lidz, J., & Christophe, A. (2019). Prosody and function words cue the acquisition of word meanings in 18-month-old infants. *Psychological Science, 30*(3), 319–332.

de Carvalho, A., Lidz, J., Tieu, L., Bleam, T., & Christophe, A., (2016). English-speaking preschoolers can use phrasal prosody for syntactic parsing. *Journal of the Acoustical Society of America, 139*(6), EL216–EL222.

DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208*(4448), 1174–1176. https://doi.org/ 10.1038/050458a0

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

Delattre, P. (1961). The intonation model of Simone de Beauvoir: A declarative comparative study on intonation. *French Review, 35*, 59–67.

Dell, F., & Halle, J. (2009). Comparing musical textsetting in French and in English songs. In J. Aroui & A. Arleo (Eds.), *Towards a Typology of Poetic Forms: From language to metrics and beyond* (pp. 63-78). John Benjamins Publishing Company. https://doi.org/10.1075/lfab.2.03del

DfE. (2011). *Modern foreign languages (MFL) in the Primary National Curriculum until 2014.* Retrieved from

https://webarchive.nationalarchives.gov.uk/ukgwa/20140107113815/http://www.education.gov.uk/schools/teachingandlearning/curriculum/primary/b00199137/mfl Accessed 11 December 2024.

DfE. (2012). *Making Foreign Languages compulsory at Key Stage 2, Consultation Report: Overview.* Retrieved from https://dera.ioe.ac.uk/id/eprint/14904/9/mfl%20compulsory%20at%20ks2%20consultation%20report.pdf Accessed 14 December 2024.

DfE. (2013a). *Languages Programmes of Study: Key Stage 2*. Crown Copyright. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-languages-progammes-of-study/national-curriculum-in-england-languages-progammes-of-study#key-stage-2-foreign-language Accessed 5 February 2023.

DfE. (2013b). *National Curriculum: Linda Dupret on Languages*. [Video] YouTube: https://youtu.be/nKPcpJtfKhI?feature=shared

DfE. (2022). The link between absence and attainment at KS2 and KS4. Crown copyright. Retrieved from https://explore-education-statistics.service.gov.uk/find-statistics/the-link-between-absence-and-attainment-at-ks2-and-ks4 Accessed 4 June 2024.

DfES. (2002). *Languages for All: Languages for Life. A Strategy for England.* Accessed on 12th December 2024: https://www.education-uk.org/documents/pdfs/2002-languages-for-all.pdf

DfES. (2005a). *KS2 Framework for Languages*. Accessed online 12th Dec 2024: https://www.lightbulblanguages.co.uk/resources/PrimaryFrench/KS2-framework-pt1.pdf

DfES. (2005b). *The Languages Ladder – Steps to Success*. Accessed on 13th December 2024: https://dera.ioe.ac.uk/id/eprint/6054/7/6page_leaflet_Redacted.pdf

*Diakou, M. (2014). *Using Songs to Enhance Language Learning and Skills in the Cypriot Primary MFL Classroom*. EdD thesis, The Open University.

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke: Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2019). How Do I Know What My Theory Predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2021). How to Use and Report Bayesian Hypothesis Tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9–26. https://doi.org/10.1037/cns0000258

Dominey, P.F., & Dodane, C. (2004). Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics, 17*(2–3), 121–145.

*Dominguez, D. (1991). *Developing language through a musical program and its effect on the reading achievement of Spanish-speaking migrant children*. Doctoral thesis, Wester Michigan University.

Doughty, C.J. (2010). Instructed SLA: constraints, compensation and enhancement. In C.J. Doughty & M.H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 256–310). Oxford, UK: Blackwell. https://doi.org/10.1002/9780470756492.ch10

Driscoll, P., & Frost, D. (Eds.). (1999). *Teaching Modern Languages in the Primary School* (1st ed.). London: Routledge. https://doi.org/10.4324/9780203983430

Driscoll, P., Jones, J., & Macrory,. G. (2004). *The Provision of Foreign Language Learning for Pupils at Key Stage 2.* London: Department for Education and Skills.

Dunn, Lloyd M., Thériault-Whalen, C.M., & Dunn, L.M. (1993). *Echelle de vocabulaire en images Peabody: adaptation française du Peabody vocabulary test-revised*. Toronto, Canada: Psycan.

Dunn, L.M., & Dunn, L.M. (1981). *Peabody Picture Vocabulary Test-Revised*. Circle Pines, MN: American Guidance Service, Inc.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning, 5*(1), 1–13.

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A destressing "deafness" in French? *Journal of memory and language, 36*(3), 406–421.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress "deafness". *The Journal of the Acoustical Society of America, 110*(3), 1606–1618.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2010). Limits on bilingualism revisited: Stress 'deafness' in simultaneous French–Spanish bilinguals. *Cognition, 114*(2), 266-275.

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress 'deafness': The case of French learners of Spanish. *Cognition, 106*(2), 682-706.

Education Act (2002). Available at https://www.legislation.gov.uk/ukpga/2002/32/contents. Accessed on 11th December 2024.

Eggermont, J.J., & Moore, J.K. (2012). Morphological and Functional Development of the Auditory Nervous System. In L. Werner, R. Fay, & A. Popper (Eds.) *Human Auditory Development. Springer Handbook of Auditory Research, vol 42*, pp. 61-105. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1421-6_3

Eimas, P.D., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in early infancy. *Science*, 171, 304–306.

Ellis, G. (2014). 'Young learners': Clarifying our terms. *ELT Journal, 68*(1), 75-78. https://doi.org/10.1093/elt/cct062

Ellis, N. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 33–68). Cambridge: Cambridge University Press.

Ellis, R. (1995). Implicit/explicit knowledge and language pedagogy, *TESOL Quarterly*, 28.

Ellis, R. (2003). *Task-based Language Learning and Teaching.* Oxford: Oxford University Press.

Endress, A.D., & Mehler, J. (2009). Primitive computations in speech processing. The *Quarterly Journal of Experimental Psychology, 62*(11). 2187–2209.

Engh, D. (2013). Why use music in English language learning? A survey of the literature. *English Language Teaching, 6*(2), 113–27. https://doi.org/10.5539/elt.v6n2p113

Erlam, R. (2006). Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study. *Applied Linguistics*, *27*(3), 464–491. https://doi.org/10.1093/applin/aml001

European Commission. (2017). *Key Data on Teaching Languages at School in Europe – 2017 Edition. Eurydice Report.* Luxembourg: Publications Office of the European Union. https://eurydice.eacea.ec.europa.eu/publications/key-data-teaching-languages-school-europe-2017-edition Accessed on 11th December 2024.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fautley, M., Kinsella, V., & Whittaker, A. (2018). *Primary School Music Teachers Survey 2018. Birmingham Music Hub*. Birmingham City University, Services for Education. Retrieved from https://bmep.servicesforeducation.co.uk/wp-content/uploads/2019/01/BMEP-Primary-Survey-2018.pdf

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In H. Papoušek, U. Jürgens, & M. Papoušek (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches*, pp.262–282. Editions de la Maison des Sciences de l'Homme; Cambridge University Press.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language, 16*(3), 477-501.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. London: Sage.

Finch, K., Theakston, A., & Serratrice, L. (2020). Teaching modern foreign languages in multilingual classrooms: an examination of Key Stage 2 teachers' experiences. *The Language Learning Journal*, *48*(5), 628–642. https://doi.org/10.1080/09571736.2018.1448432

Fisher, R.A. (1935). *The Design of Experiments*. Tweeddale: Oliver and Boyd.

Fonseca-Mora, M.C. (2000). Foreign language acquisition and melody singing. *ELT Journal, 54*(2), 146–152. http://dx.doi.org/10.1093/elt/54.2.146

Fonseca-Mora, M.C., & Gant, M. (2016). *Melodies, Rhythm and Cognition in Foreign Language Learning* (M. C. Fonseca-Mora & M. Gant, Eds.). Newcastle, UK: Cambridge Scholars Publishing.

*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

Forder, C., Phillips, H., & Watts, C. (2013). *Living languages: an integrated approach to teaching foreign languages in primary schools.* UK: Routledge. https://doi.org/10.4324/9780203809396

Formby, D. (1967). Maternal recognition of infant's cry. *Developmental Medicine and Child Neurology*, 9, 293–298.

Forster, E. (2006). The value of songs and chants for young learners. *Encuentro*, *16*, 63–68.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T.H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews, 92*(4), 1941–1968. https://Doi.Org/10.1111/Brv.12315

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. Https://Doi.Org/10.1126/Science.1255484

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

François, C., Teixidó, M., Takerkart, S., Agut, T., Bosch, L., & Rodriguez-Fornells, A. (2017). Enhanced neonatal brain responses to sung streams predict vocabulary outcomes by age 18 months. *Scientific Reports, 7*(1), 12451. doi: 10.1038/s41598-017-12798-2

Frank, M.C., Alcock, K.J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S., Bergelson, E., Bergmann, C., Black, A.K., Blything, R., Böhland, M.P., Bolitho, P., Borovsky, A., Brady, S.M., Braun, B., Brown, A., Byers-Heinlein, K., Campbell, L.E., … Soderstrom, M. (2020). Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Fukui, N. (2006). Phrase structure. In N. Fukui (Ed.) *Theoretical Comparative Syntax: Studies in Macroparameters* (1st ed.), pp.258–288. London: Routledge. https://doi.org/10.4324/9780203479179

Gaillard, S., & Tremblay, A. (2016). Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. *Language Learning*, *66*(2), 419–447. https://doi.org/10.1111/lang.12157

Gallagher, J. (2019). *Learning Languages in Early Modern England*. Oxford: Oxford University Press.

Gamble, C.J., & Smalley, A. (1975). Primary French in the Balance: Were the Scales Accurate? *Modern Languages: Journal of the Modern Language Association*, *56*(2), 94–97.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. NY, USA: Basic Books.

Garnica, O.K. (1977). Some prosodic and paralinguistic features of speech to young children. In C. Snow & C.A. Ferguson (Eds.), *Talking to children: Language input and acquisition*, pp.63–88. Cambridge: Cambridge University Press.

Garton, S., Copland, F., & Burns, A. (2011). *Investigating Global Practices in Teaching English to Young Learners*. London: British Council.

Gass, S. (2019). A WORD FROM THE EDITOR. *Studies in Second Language Acquisition, 41*(1), 1–2. doi:10.1017/S0272263119000081

Gathercole, S. & Baddeley, A. (1993). *Working Memory and Language*. Hove, UK: Lawrence Erlbaum Associates.

Geisler, P. (2008). *Musikorientiertes Lernen im Englisch-Unterricht der Grundschule*. Doctoral thesis, Pädagogische Hochschule Freiburg.

Gerken, L., Jusczyk, P.W., & Mandel, D.R. (1994). When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition, 51*(3), 237–265. https://doi.org/10.1016/0010-0277(94)90055-8

Gervain, J. (2018). The role of prenatal experience in language development. *Current Opinion in Behavioral Sciences*, 21, 62–67. https://doi.org/ 10.1016/j.cobeha.2018.02.004

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: Oxford University Press.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., and Mehler, J. (2008). Bootstrapping word order in prelexical infants: a Japanese-Italian crosslinguistic study. *Cogn. Psychol.* 57, 56–74.

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.

Gleitman, L.R. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L.R. Gleitman (eds.), *Language acquisition: The state of the art*, pp.3–48. Cambridge, MA: Cambridge University Press.

*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

Gleitman, L., Gleitman, H., Landau, B., & Wanner, E. (1987). Where learning begins: Initial representations for language learning. In E. Newmayer (Ed.), *The Cambridge Linguistic Survey (Vol. 11)*, pp.150–193. New York: Cambridge University Press.

Goad, H., & White, L. (2004). Ultimate attainment of L2 inflection: Effects of L1 prosodic structure. In S. Foster-Cohen, M. Sharwood Smith, A. Sorace, & M. Ota (Eds.), *EUROSLA Yearbook 4*, pp.119–145. Amsterdam: John Benjamins Publishing.

Goad, H., & White, L. (2006). Ultimate attainment in interlanguage grammars: A prosodic approach. *Second Language Research*, 22, 243–268. https://doi.org/10.1191/0267658306sr268oa

Goad, H., & White, L. (2009). Prosodic transfer and the representation of determiners in Turkish-English interlanguage. In N. Snape, Y.-K. I. Leung, & M. Sharwood-Smith (Eds.), *Representational deficits in SLA: Studies in honor of Roger Hawkins*, pp.1–26. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/lald.47.04goa

Goad, H. & White, L. (2019). Prosodic effects on L2 grammars. *Linguistic Approaches to Bilingualism, 9*(6), 769–808. https://doi.org/10.1075/lab.19043.goa

Gorard, S. (2001). *Quantitative methods in educational research: the role of numbers made easy.* London: Continuum.

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.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, 110, 47–59.

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.

Gout, A., Christophe, A. & Morgan, J.L. (2004). Phonological phrase boundaries constrain lexical access. II. Infant data. *Journal of Memory and Language 51*(4). 548–567.

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early Language Learning: The Impact of Teaching and Teacher Factors. *Language Learning*, *67*(4), 922–958. https://doi.org/10.1111/lang.12251

Granier-Deferre C., Bassereau S., Ribeiro A., Jacquet, A-Y., DeCasper, A.J. (2011). A Melodic Contour Repeatedly Experienced by Human Near-Term Fetuses Elicits a Profound Cardiac Reaction One Month after Birth. *PLoS ONE, 6*(2): e17304. doi:10.1371/journal.pone.0017304

Grebe, K., Grebe, W. (1975). Verb tone patterns in Lamnsok. *Linguistics, 149*, 5–24.

Green, J.A., & Gustafson, G.E. (1983). Individual recognition of human infants on the basis of cries alone. *Developmental Psychobiology*, 16, 485–493.

Grieser, D.L., & Kuhl, P.K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology, 24*(1), 14–20. https://doi.org/10.1037/0012-1649.24.1.14

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker & J.J. McCarthy (Eds.), *The logical problem of language acquisition*, pp.183–210. Cambridge, MA: MIT Press.

Guardian (2002). *Clarke targets primary schools in language strategy.* Accessed on 12th December 2024: https://www.theguardian.com/education/2002/dec/18/schools.highereducation

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

Gussenhoven, C., & Chen, A. (Eds.). (2020). *The Oxford Handbook of Language Prosody*. Oxford: Oxford University Press.

Gustafson, G.E., Sanborn, S.M., Lin, H. & Green, J.A. (2017). Newborns' Cries are Unique to Individuals (But Not to Language Environment). *Infancy, 22*(6), 736–747. https://doi.org/10.1111/infa.12192

Gustafsson, E.E., Levrero, F., Reby, D., & Mathevon, N. (2013). Fathers are just as good as mothers at recognizing the cries of their baby. *Nature Communications*, 4, 1–6.

Gutman, A., Dautriche, I., Crabbé, B., & Christophe, A. (2015). Bootstrapping the Syntactic Bootstrapper: Probabilistic Labeling of Prosodic Phrases. *Language Acquisition, 22*(3), 285–309. https://doi.org/10.1080/10489223.2014.971956

*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.

Hallé, P.A., Durand, C., & de Boysson-Bardies, B. (2008). Do 11-month-old French infants process articles? *Language and Speech, 51*(1–2). 23–44.

Hamilton, C., Chalmers, H., & Murphy, V.A. (2024, January 24). *Investigating the efficacy of whole-class singing activities for linguistic outcomes of young language learners in English primary schools.* Retrieved from osf.io/yp43w

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, 52*(8), 1488–1509. https://doi.org/10.1080/03004279.2023.2168132

Hamilton, C., Schulz, J., Chalmers, H., & Murphy, V. (2024). 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. *System, 124*(103350).

Hardach, S. (15th April 2020). Do Babies Cry in Different Languages? A pioneering German researcher decodes newborns' cries. Here's what they reveal. *New York Times*. Accessed on 12th January 2025 at https://www.nytimes.com/2020/04/15/parenting/baby/wermke-prespeech-development-wurzburg.html

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:

Nikolov, M. (ed.) *Early Learning of Modern Foreign Languages. Processes and outcomes.* Bristol, Buffalo, Toronto: Multilingual Matters, pp.1–14.

Hauser, M.D., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*, B53–B64.

Hauser, M.D., Weiss, D. & Marcus, G. (2002). Rule learning by cotton-top tamarins. *Cognition, 86*, B15–B22.

Hawkes, R. (2024). *Language Progression: French KS2 Curriculum*. Accessed on 17th December 2024: https://www.rachelhawkes.com/Resources/PrFrench/Yr34Autumn.php

Hawkins, E. (1984). *Awareness of Language: An Introduction*. Cambridge: Cambridge University Press.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic inquiry, 20*(2), 253–306.

Hayes, B. (1995). *Metrical stress theory: principles and case studies.* Chicago: University of Chicago Press.

Heinen, K.S., & Kadow, H. (1990). The Acquisition of French by Monolingual Children: A Review of the Literature. In J.M. Meisel (Ed.). *Two first languages: Early grammatical development in bilingual children*. (pp.47–72). Dordrecht, Netherlands: Foris Publications.

*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

Hirschberg, J., Beňuš, Š., Gravano, A., & Levitan, R. (2020). Prosody in discourse and speaker state. In C. Gussenhoven & A. Chen (Eds.). *The Oxford Handbook of Language Prosody*. pp.468–476. Oxford: Oxford University Press.

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics, 47*(2), 359–382. https://doi.org/10.1515/ling.2009.013

Holmes, B., & Myles, F. (2019). *White Paper: Primary Languages Policy in England – The Way Forward*. RiPL. Retrieved from www.ripl.uk/policy/ Accessed 10 November 2021.

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 May 2, 2022

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

Hsu, H.C., Fogel, A., & Cooper, R.B. (2000). Infant vocal development during the first 6 months: Speech quality and melodic complexity. *Infant and Child Development: An International Journal of Research and Practice, 9*(1), 1–16.

Hunt, M., Barnes, A., Powell, B., Lindsay, G. & Muijs, D. (2005). Primary modern foreign languages: an overview of recent research, key issues and challenges for educational policy and practice. *Research Papers in Education, 20*(4), 371–390. https://doi.org/10.1080/02671520500335774

Hunt, M. (2009). Progression and assessment in foreign languages at Key Stage 2. *The Language Learning Journal, 37*(2), 205–217. https://doi.org/10.1080/09571730902928086

Hyams, N. (1988). A principles-and-parameters approach to the study of child language. *Papers and Reports of Child Language Development*, 27, 153–161.

Hymes, D. (1972). On Communicative Competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp.269–293). Harmondsworth: Penguin Books.

Ireland, K., Parker, A., Foster, N., & Penhune, V. (2018). Rhythm and Melody Tasks for School-Aged Children With and Without Musical Training: Age-Equivalent Scores and Reliability. *Frontiers in Psychology*, *9*, 426. https://doi.org/10.3389/fpsyg.2018.00426

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

Isbell, D. (2024). Open science, data analysis, and data sharing. In L. Plonsky (Ed.), *Open science in applied linguistics* (pp. 104-122). Applied Linguistics Press.

Isbell, D., Brown, D., Chen, M., Derrick, D., Ghanem, R., Gutiérrez Arvizu, M.N., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *Modern Language Journal*, 106, 172-195.

Jaekel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From Early Starters to Late Finishers? A Longitudinal Study of Early Foreign Language Learning in School. *Language Learning*, *67*(3), 631–664. https://doi.org/10.1111/lang.12242

*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

Jeffreys, H. (1961). *The Theory of Probability. 3rd edition.* Oxford: Oxford University Press.

*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*"], 15*(2), 57–75.

Jiang, R. (2025). *The effect of orthographic input on young Mandarin-speaking EFL children's English pronunciation learning.* Doctoral thesis, University of Oxford.

John, E. (1973). Saturday night's alright for fighting [Song]. On *Goodbye yellow brick road*. DJM.

Johnson, E.K., Seidl, A., & Tyler, M.D. (2014). The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PloS one, 9*(1), e83546.

Jolly, Y. (1975). The Use of Songs in Teaching Foreign Languages. *The Modern Language Journal, 59*(1/2), 11–14.

Jones, J., & Coffey, S. (2016). *Modern Foreign Languages 5-11: A guide for teachers* (3rd ed.). Routledge. https://doi.org/10.4324/9781315628028

Jongman, A. & Tremblay, A. (2020). Word prosody in second language acquisition. In C. Gussenhoven & A. Chen (Eds.) *The Oxford Handbook of Language Prosody.* pp.594–604. Oxford: Oxford University Press.

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

Jusczyk, P.W. (1989). *Perception of Cues to Clausal Units in Native and Non-Native Languages.* Paper presented at the biennial meeting of the Society for Research in Child Development, Kansas City, Mo., April 1989.

Jusczyk, P.W. (1997). *The Discovery of Spoken Language*. Cambridge, M.A.: MIT Press.

Jusczyk, P.W., Hohne, E. & Mandel, D.R. (1995). Picking up regularities in the sound structure of the native language. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research*, pp.91–119. Baltimore: York Press.

Jusczyk, P.W., Luce, P.A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630–645.

Jusczyk, P.W., Mazuka, R., Mandel, D., Kiritani, S., & Hayashi, A. (1993). *A cross-linguistic study of American and Japanese infants' perception of acoustic correlates to clausal units*. In biennial Meeting of the Society for Research in Child Development, New Orleans, La, March 1993.

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

Kedar, Y., Casasola, M., & Lust, B. (2006). Getting there faster: 18-and 24-month-old infants' use of function words to determine reference. *Child Development, 77*(2). 325–338.

Kelly, L.G. (1976). *25 Centuries of Language Teaching: 500BC – 1969. 2nd Edition.* Massachusetts, USA: Newbury House Publishers.

Kemler Nelson, D.G., Hirsh-Pasek, K., Jusczyk, P.W., & Wright-Cassidy, K. (1989). How prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16, 55–68.

Kent, R.D., & Murray, A.D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *The Journal of the Acoustical Society of America, 72*(2), 353–365.

*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.

Kiparsky, P. (1977). The rhythmic structure of English verse. *Linguistic Inquiry, 8*(2), 189-247.

Kiparsky, P. (2020). Stress, meter, and text-setting. In C. Gussenhoven & A. Chen (Eds.). *The Oxford Handbook of Language Prosody*. pp.657–675. Oxford: Oxford University Press.

Kirsch, C. (2008). *Teaching Foreign Languages in the Primary School.* London: Continuum.

Kisilevsky, B.S., Hains, S.M.J., Brown, C.A., Lee, C.T., Cowperthwaite, B., Stutzman, S.S., Swansburg, M.L., Lee, K., Xie, X., Huang, H., Ye, H.H., Zhang, K., & Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development, 32*(1), 59–71. https://doi.org/10.1016/j.infbeh.2008.10.002

Klein, M., Sosu, E.M., & Dare, S. (2022). School Absenteeism and Academic Achievement: Does the Reason for Absence Matter? *AERA Open*, 8. https://doi.org/10.1177/23328584211071115

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

knoxstudy (2023, May 23) [social media post] Newborn babies have accents! Knox Study OGs may remember this from 2020. #language #accents #interesting #todayilearned. *Instagram*. Accessed on 12th January 2025 at https://www.instagram.com/reel/CsmE29Jol1B/?igshid=MmJiY2I4NDBkZg%3D%3D

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.

Krashen, S., & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. Hayward, CA: The Alemany Press.

Kreiner, H., & Eviatar, Z. (2014). The missing link in the embodiment of syntax: prosody. *Brain and Language*, 137, 91–102. https://doi.org/10.1016/j.bandl.2014.08.004

Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831–843.

Kutner, M.H. (2005). *Applied linear statistical models*. Boston: McGraw-Hill Irwin.

Ladefoged, P. (1975). *A Course in Phonetics.* New York: Harcourt Brace Jovanovich.

Lanvers, U. (2017). Contradictory Others and the Habitus of Languages: Surveying the L2 Motivation Landscape in the United Kingdom. *The Modern Language Journal*, *101*(3), 517–532. https://doi.org/10.1111/modl.12410

Leach, E.E. (2005). Learning French by singing in 14th-century England. *Early Music, 33*(2), 253–272. https://doi.org/10.1093/em/cah069

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

*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

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachusetts: MIT Press.

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

Léopold, W., Jones, R., Ervin-Tripp, S., Rivers, W. & Malherbe, E. (1969). 1. How and when do persons become bilingual? Comment et quand devient-on bilingue?. In L. Kelly (Ed.), *Description and Measurement of Bilingualism: An International Seminar, University of Moncton June 6-14, 1967* (pp. 11–78). Toronto: University of Toronto Press. https://doi.org/10.3138/9781487589134-005

*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

Levine, D., Hirsh-Pasek, K., & Golinkoff, R.M. (2020). Infant Word Learning and Emerging Syntax. In J.J. Lockman & C.S. Tamis-LeMonda (Eds.), *The Cambridge Handbook of Infant Development: Brain, Behavior, and Cultural Context* (pp. 632–660). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108351959.023

Lewkowicz, D.J., & Hansen-Tift, A.M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech, Proc. Natl. Acad. Sci. U.S.A., 109(5), 1431–1436, https://doi.org/10.1073/pnas.1114783109

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry, 8*(2), 249–336.

Liu, M. (2023). Whose open science are we talking about? From open science in psychology to open science in applied linguistics. *Language Teaching, 56*(4), 443–450. doi:10.1017/S0261444823000307

Lightbown, P.M., & Spada, N. (2003). *How Languages are Learned.* Oxford: Oxford University Press.

Light Bulb Languages. (2025). *Continuing and developing KS2 pedagogy in KS3*. Accessed on 16th December 2024: https://www.lightbulblanguages.co.uk/transition-KS2pedagogy.htm

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

Long, M.H. (1997). *Focus on Form in Task-based Language Teaching.* Manoa: University of Hawaii.

Long, M.H. (2015). *Second language acquisition and task-based language teaching (1st ed.)*. UK: John Wiley & Sons.

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

LostTheGameOfThrones. (2022). *Of all the schools I've worked at, only one has ever had a specialist teacher for MFL; even then, they…* [Comment on the online forum post, *Thoughts on the teaching of MFL in primary*]. Reddit. Accessed on 16th December 2024: https://www.reddit.com/r/TeachingUK/comments/xznhe4/comment/irn4ge6/

*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*. Doctoral thesis, University of Illinois at Urbana-Champaign.

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

*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.

Luscombe, D. (2004). Thought and Learning. In D. Luscombe & J. Riley-Smith (Eds.), *The New Cambridge Medieval History* (pp. 461–498). Cambridge: Cambridge University Press.

*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.

Macaro, E. (2008). The decline in language learning in England: getting the facts right and getting real. *The Language Learning Journal*, *36*(1), 101–108. https://doi.org/10.1080/09571730801988595

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language, 31*(4), 883–914. https://doi.org/10.1017/s0305000904006336

*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

Mampe, B., Friederici, A.D., Christophe, A. & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Curr. Biol*. 19, 1994–1997. https://doi.org/10.1016/j.cub.2009.09.064

Mandel, D.R., Jusczyk, P.W., Mazuka, R., Kiritane, S., & Hayashi, A. (1992). *Perception of Japanese clauses by American 4$_{1/2}$ month olds.* In Workshop on cross-language speech perception, Tampa, Fla.

Mandel, D.R., Kemler Nelson, D.G., & Jusczyk, P.W. (1996). Infants remember the order of words in a spoken sentence. *Cogn. Dev.*, 11, 181–196.

Marsden, E., Mackey A., & Plonsky, L. (2015). The IRIS Repository of Instruments for Research into Second Languages: Advancing methodology and practice. In Mackey, A. & Marsden E. (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages*. Chapter 1. New York: Routledge.

Martin, C. (2000). Modern foreign languages at primary school: a three-pronged approach? *The Language Learning Journal, 22*(1), 5–10. https://doi.org/10.1080/09571730085200181

Martin, J.A.M. (1981). *Voice, speech, and language in the child: Development and disorder (Vol. 4).* Springer Science & Business Media.

Martinez-Alvarez, A., Benavides-Varela, S., Lapillonne, A., & Gervain, J. (2023). Newborns discriminate utterance-level prosodic contours. *Developmental Science, 26*(2), e13304. https://doi.org/10.1111/desc.13304

Matthews, P.H. (2014). *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford University Press.

Mayberry, R.I., & Squires, B. (2006). Sign language: acquisition. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed.), 11, 291–296. Oxford: Elsevier.

Maye, J., Werker, J.F., & Gerken, L.A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–B111.

*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.

McCoy C.E. (2017). Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *The Western Journal of Emergency Medicine, 18*(6), 1075–1078. https://doi.org/10.5811/westjem.2017.8.35985

McDade, H.L., Simpson, M.A., & Lamb, D.E. (1982). The use of elicited imitation as a measure of expressive grammar: a question of validity. *Journal of Speech and Hearing Disorders*, *47*(1), 19–24.

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

Mehler, J., Jusczyk, P.W., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, 143–178.

Mende, W., Wermke, K., Schindler, S., Wilzopolski, K., & Hoeck, S. (1990). Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders. *Early Child Development and Care*, 65, 95–107.

Míguez, M. (2017). *Ways to use song lyrics to improve comprehension.* [Blog, 5th July]. British Council. Accessed 20th December 2024: https://www.britishcouncil.org/voices-magazine/ways-use-song-lyrics-improve-comprehension

Millotte, S., Morgan, J., Margules, S., Bernal, S., Dutat, M., & Christophe, A. (2011). Phrasal prosody constrains word segmentation in French 16-month-olds. *Journal of Portuguese Linguistics, 10*(1), 67–86. https://doi.org/10.5334/jpl.101

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

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the Quality of Randomized Controlled Trials: An Annotated Bibliography of Scales and Checklists. *Controlled Clincial Trials*, 16, 62–73.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, 339. https://doi.org/10.1136/bmj.b2535

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

Moon, C., Cooper, R.P., & Fifer, W.P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*(4), 495–500. https://doi.org/10.1016/0163-6383(93)80007-U

Moon, C., Lagercrantz, H., & Kuhl, P.K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Paediatr*, 102, 156-160. https://doi.org/10.1111/apa.12098

*Moradi, F., & Shahrokhi, M. (2014). The effect of listening to music on Iranian children's segmental and suprasegmental pronunciation. *English Language Teaching, 7*(6): 128–42.

Mordsley, J. (2017). *Why use rhythm, rhyme and repetition in class?* [Blog, 4th October]. British Council. Accessed on 20th December 2024: https://www.britishcouncil.org/voices-magazine/why-use-rhythm-rhyme-and-repetition-language-class

Morgan, J.L. (1986). *From simple input to complex grammar.* Cambridge: MIT Press.

Morgan, J.L. (1990). Input, innateness, and induction in language acquisition. *Developmental Psychobiology, 23*(7), 661–678. https://doi.org/10.1002/dev.420230709

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

Muijs, D., Barnes, A., Hunt, M., Powell, B., Arweck, E., Lindsay, G. & Martin, C. (2005). *Evaluation of the Key Stage 2 Language Learning Pathfinders: Department for Education and Skills Research Report 692*. Nottingham: DfES Publications. http://www.dfes.gov.uk/research/data/uploadfiles/RR692.pdf

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks; what they measure and how they relate to each other. In E.E. Tarone, S. Gass, & A.D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 227–243). Hove: Lawrence Erlbaum.

Muñoz, C. (2006). *Age and the rate of foreign language learning.* Bristol, UK: Multilingual Matters.

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

Murphy, J. (1968). Religion, the State, and Education in England. *History of Education Quarterly, 8*(1), 3–34.

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

Murphy, V.A., Macaro, E., Alba, S. & Cipolla, C. (2015). The influence of learning a second language in primary school on developing first language literacy skills. *Applied Psycholinguistics, 36*(5), 1133–1153. https://doi.org/10.1017/s0142716414000095

*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)*, 512–517. https://doi.org/10.2991/aes-18.2019.115

Myles, F. (2017). Learning foreign languages in primary schools: is younger better? *Languages, Society & Policy*. https://doi.org/10.17863/CAM.9806

Nation, I.S.P. & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.). *Handbook of Research in Second Language Teaching and Learning*, Volume III, Chapter 26. New York: Routledge.

*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*(1), 108–121.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998a). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756–66. https://doi.org/10.1037//0096-1523.24.3.756

Nazzi, T., Floccia, C., Bertoncini, J. (1998b). Discrimination of pitch contours by neonates. *Infant Behav. Dev., 21*(4), 779–784.

Nespor, M., & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

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

Newport, E.L., Hauser, M.D., Spaepen, G., & Aslin, R.N. (2004). Learning at a distance: II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *49*, 85–119.

Ní Chasaide, A., & Gobl, C. (2004). Voice quality and f0 in prosody: towards a holistic account. *Proc. Speech Prosody 2004*, 189–196, doi: 10.21437/SpeechProsody.2004-44

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science, 7*(6), 615–631. Https://Doi.Org/10.1177/1745691612459058

Nunan, D. (2004). *Task-Based Language Teaching.* Cambridge: Cambridge University Press.

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

Office for National Statistics (ONS; 2021). *Census 2021.* Available at: https://www.ons.gov.uk/census Accessed 10 June 2024.

Ommen, S. van, Boll-Avetisyan, N., Larraza, S., Wellmann, C., Bijeljac-Babic, R., Höhle, B. & Nazzi, T. (2020). Language-specific prosodic acquisition: A comparison of phrase boundary perception by French- and German-learning infants. *Journal of Memory and Language*, 112, 104108. https://doi.org/10.1016/j.jml.2020.104108

Open Science Framework, OSF (n.d.) *Preregistration*. https://help.osf.io/article/145-preregistration

Orme, N. (2006). *Medieval schools: from Roman Britain to Renaissance England*. USA: Yale University Press.

Outters, V., Schreiner, M.S., Behne, T., & Mani, N. (2020). Maternal input and infants' response to infant-directed speech. *Infancy, 25*(4), 478-499.

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., Aki, 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. *PLOS Medicine, 18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Papoušek, M., & Hwang, S.-F.C. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics, 12*(4), 481–504. doi:10.1017/S0142716400005889

Papoušek, M., Papoušek, H., & Haekel, M. (1987). Didactic adjustments in fathers' and mothers' speech to their 3-month-old infants. *Journal of Psycholinguistic Research, 16*(5), 491–516.

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

Paradis, J., Nicoladis, E., & Genesee, F. (2000). Early emergence of structural constraints on code-mixing: Evidence from French-English bilingual children. *Bilingualism: Language and Cognition, 3*(3), 245–261.

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

Partanen, E., Kujala, T., Näätänen, R., Liitola, A., Sambeth, A., Huotilainen, M. (2013). Learning-induced neural plasticity of speech processing before birth. *Proc. Natl. Acad. Sci. USA, 110*(37), 15145–15150. https://doi.org/10.1073/pnas.1302159110

Pearson (1999). *Weschler Abbreviated Scale of Intelligence (WASI).* USA: Pearson.

Peters, A. (1983). *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Peters, A. (1985). Language segmentation: Operating principles for the perception and analysis of language. In D.I. Slobin (Ed.). *The Crosslinguistic Study of Language Acquisition: Volume 2: Theoretical Issues (1st ed.)*, pp.1029–1067. New York: Psychology Press.

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

Phakiti, A. (2014). *Experimental Research Methods in Language Learning*. London: Bloomsbury.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation.* Doctoral thesis, Massachusetts Institute of Technology.

Pike, K. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Pilon, R. (1981). Segmentation of speech in a foreign language. *Journal of Psycholinguistic Research*, 10, 113–121.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Plonsky, L. (Ed.) (2024a). Open science in applied linguistics. Applied Linguistics Press.

Plonsky, L. (2024b). The era of open science is upon us (Or, why a more open science is also a higher quality science). In L. Plonsky (Ed.), *Open science in applied linguistics* (pp. 7-16). Applied Linguistics Press.

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), 529–46.

Porte, G., & McManus, K. (2019). *Doing Replication Research in Applied Linguistics.* London: Routledge.

Post, B., & Payne, E. (2018). Speech rhythm in development: What is the child acquiring? In P. Prieto & N. Esteve-Gibert (Eds.) *The development of prosody in first language acquisition* (pp.125–143). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.23.07pos

Powell, B., Wray, D., Rixon, S., Medwell, J., Barnes, A., & Hunt, M. (2000). *Analysis and evaluation of the current situation relating to the teaching of Modern Foreign Languages at Key Stage 2 in England: Research report commissioned by the Qualifications and Curriculum Authority*. Coventry: University of Warwick. Accessed 22nd December 2024: https://warwick.ac.uk/fac/soc/ces/research/teachingandlearning/resactivities/subjects/mfl/primarylanguages/qca_lang_report.pdf

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America, 90*(6), 2956–2970. https://doi.org/10.1121/1.401770

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

Prieto, P., & Esteve-Gibert, N. (2018). *The development of prosody in first language acquisition.* Amsterdam: John Benjamins Publishing Company.

Prochnow, A., Erlandsson, S., Hesse, V., & Wermke, K. (2019). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Musicae Scientiae, 23*(2), 143-156. https://doi.org/10.1177/1029864917733035

Pujol, R., Lavigne-Rebillard, M., & Uziel, A. (1991). Development of the Human Cochlea. *Acta Oto-Laryngologica, 111*(sup482), 7–13. https://doi.org/10.3109/00016489109128023

QCA; Qualifications and Curriculum Authority. (2007). MFL French at key stage 2. Retrieved from

https://webarchive.nationalarchives.gov.uk/ukgwa/20090608173756/http://www.standards.dfes.gov.uk/schemes3/subjects/primary_mff/?view=get Accessed on 20th December 2024.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265–292.

Ramus, F., Hauser, M., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*(5464), 349–351. doi: 10.1126/science.288.5464.349

Räsänen, O., Altosaar, T., & Laine, U.K. (2008). Comparison of prosodic features in Swedish and Finnish IDS/ADS speech. *Proc. of Nordic Prosody X.*

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

Reynolds, C. R., & Kamphaus, R. W. (2015). *RIAS-2: Reynolds Intellectual Assessment Scales, Second Edition*. Lutz, FL: Psychological Assessment Resources.

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.

Rixon, S. (1991). The role of fun and games activities in teaching young learners. In *Teaching English to Children: From Practice to Principle*. Eds. C. Brumfit, Moon, J., & Tongue, R. (pp.33–48). London: Collins.

Robertson, S., von Hapsburg, D., & Hay, J.S. (2013). The effect of hearing loss on the perception of infant-and adult-directed speech. *Journal of Speech, Language, and Hearing Research, 56*(4), 1108–1119.

Robertson, C. & Salter, W. 2007. *The Phonological Awareness Test 2* [Measurement Instrument]. Austin, TX: PRO-ED, Inc.

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*, 100436. https://doi.org/10.1016/j.edurev.2022.100436

Rowland, C. (2014). *Understanding Child Language Acquisition.* Abingdon: Routledge.

RStudio Team (2023). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Rumley, G. (1999). Games and songs for teaching modern foreign languages to young children. In P. Driscoll & D. Frost (Eds). *Teaching modern languages in the primary school*. pp.115–126. London: Routledge.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27–52.

Saksida, A., Flo, A., Guedes, B., Nespor, M., & Garay, M.P. (2021). Prosody facilitates learning the word order in a new language. *Cognition*, *213*, 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.

Schöpfel, J. (2010). Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues.* Prague, 6-7 December 2010, Czech Republic. pp.11–26.

Schulz, J. (2024). *Multi-word-constructions and linguistic development in early foreign languages classrooms: the role of input variability.* Doctoral thesis, University of Oxford.

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

Schulz, K.F., Altman, D.G., Moher, D., for the CONSORT Group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Lancet*. 375:9721, 1136.

*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

Seccombe, C. (2021). *Primary Languages in England, a timeline: 2002–2021*. [Infographic]. Accessed 18th December 2024: https://changing-phase.blogspot.com/2021/11/all-primary-languages-conference-online.html

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, Massachusetts: The MIT Press.

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.

Sharman, K.M., Meissel, K., Tait, J.E., Rudd, G., & Henderson, A.M. (2023). The effects of live parental infant-directed singing on infants, parents, and the parent-infant dyad: A systematic review of the literature. *Infant Behavior and Development*, 72, 101859.

Sharpe, K. (1989). Smothered Tongue. *Times Educational Supplement*, 2 June.

Sharpe, K. (1991). Primary French: More phoenix than dodo now. *Education 3-13, 19*(1), 49–53. https://doi.org/10.1080/03004279185200101

Sharpe, K. (1995). The primacy of pedagogy in the early teaching of Modern Languages. *The Language Learning Journal, 12*(1), 40–42. https://doi.org/10.1080/09571739585200411

Shattuck-Hufnagel, S. (2020). The role of phrase-level prosody in speech production planning. In C. Gussenhoven & A. Chen (Eds.). *The Oxford Handbook of Language Prosody*. pp.522–538. Oxford: Oxford University Press.

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

Shi, R., Cutler, A., Werker, J.F., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America, 119*(6). 61–67.

Shi, R., Werker, J.F., Morgan, J.L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*(2), B11–B21.

Shippey, T. (2007). I lerne song. *London Review of Books.* Accessed 10th Dec 2024 at https://www.lrb.co.uk/the-paper/v29/n04/tom-shippey/i-lerne-song

*Siebring, M.F. (2004). *The effectiveness of a systematic approach based on songs to prevent and correct errors in elementary core French*. Master's thesis, Nipissing University.

Silvey, C., Dienes, Z., & Wonnacott, E. (2021, December 3). *Bayes factors for logistic (mixed effect) models*. https://doi.org/10.31234/osf.io/m4hju

Simpson, A.J. (2015). *How to use songs in the English language classroom.* [Blog, 4th March]. British Council. Accessed 20th December 2024: https://www.britishcouncil.org/voices-magazine/how-use-songs-english-language-classroom

Siraj-Blatchford, I. (1999). Early Childhood Pedagogy: Practice, Principles and Research. In P. Mortimore (Ed.) *Understanding Pedagogy and its Impact on Learning*, pp.20–45. London: Paul Chapman.

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

Smith, R., Liddicoat, T., Consoli, S., McConachy, T., Pinter, A., Rixon, S., Sharpling, N., Smith-Dennis, E., & White, N. (2021). *Timeline of MLT in England*. University of Warwick. Accessed 10th December 2024: https://warwick.ac.uk/fac/soc/al/research/groups/llta/mlt/timeline/

Smith, S. (n.d.). *Using music in the languages classroom*. Website: https://www.frenchteacher.net/teachers-guide/using-music-in-mfl-lessons/ Accessed 17th December 2024.

Snow, C., & Ferguson, C.A. (Eds.). (1977). *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

Sopoci, McKenzie K. (2023). *The Construct Validity of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) and the Reynolds Intellectual Assessment Scales, Second Edition (RIAS-2).* Master's thesis, Eastern Illinois University. https://thekeep.eiu.edu/theses/4981 Accessed 11 June 2024.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, 1, 1–33. https://doi.org/10.1075/lab.1.1.01sor

Spinelli, M., Fasolo, M., & Mesman, J. (2017). Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review*, 44, 1–18.

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.

Stern, D.N., Spieker, S., Barnett, R.K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of child language, 10*(1), 1–15.

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

Streeter, L.A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64, 1582–1592.

Swerts, M. & Krahmer, E. (2020). Visual prosody across cultures. In C. Gussenhoven & A. Chen (Eds.). *The Oxford Handbook of Language Prosody*. pp.477–485. Oxford: Oxford University Press.

Székely, É., Henter, G.E., Beskow, J. & Gustafson, J. (2020). Breathing and Speech Planning in Spontaneous Speech Synthesis. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain,* pp.7649–7653, doi: 10.1109/ICASSP40776.2020.9054107.

Tellier, A (2019). *Primary French in the Balance – a critique of the Nuffield pilot*. RiPL Summary of points from Bennett, S. N. (1975); Buckby, M. (1976); Gamble, C. J., & Smalley, A. (1975); Kunkle, J.F. (1977).

Tenenbaum, E.J., Sobel, D.M., Sheinkopf, S.J., Malle, B.F., & Morgan, J.L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language, 42*(6), 1173–1190. doi:10.1017/S0305000914000725

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.

Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: Use of statistical and stress cues to word boundaries by 7- and 9-month-old infants. *Developmental Psychology*, 39, 706–716.

Thiessen, E.D., & Saffran, J.R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning & Development*, 3, 73–100.

Tinsley, T. (2013). *Languages: State of the Nation*. London: British Academy. https://www.thebritishacademy.ac.uk/documents/2601/Languages-state-of-the-nation-demand-supply-language-skills-UK-2013.pdf

Tinsley, T., & Doležal, N. (2018). *Language Trends 2018: Language Teaching in Primary and Secondary Schools in England Survey Report*. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/language_trends_2018_report.pdf

Tomasello, M. (1992). The social bases of language acquisition. *Social Development, 1*, 67–87.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, Massachusetts: Harvard University Press.

*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., & Nakata, T. (2001). Emotion and music in infancy. *Musicae scientiae, 5*(1_suppl), 37–61.

Trehub, S.E., Unyk, A.M. & Trainor, L.J. (1993a). Maternal singing in cross-cultural perspective. *Infant Behavior and Development,* 16, 285–295.

Trehub, S.E., Unyk, A.M., & Trainor, L.J. (1993b). Adults identify infant-directed music across cultures. *Infant Behavior and Development*, 16, 193–211.

Trouvain, J. & Braun, B. (2020). Sentence prosody in a second language. In C. Gussenhoven & A. Chen (Eds.) *The Oxford Handbook of Language Prosody*, pp.605–618. Oxford: Oxford University Press.

Tryfon A., Foster N.E., Ouimet T., Doyle-Thomas K., Anagnostou E., Sharda M., et al. (2017). Auditory-motor rhythm synchronization in children with autism spectrum disorder. *Res. Autism Spectrum Disord.* 35, 51–61. https://doi.org/10.1016/j.rasd.2016.12.004

Tsang, C.D., Falk, S., & Hessel, A. (2017). Infants prefer infant-directed song over speech. *Child development, 88*(4), 1207–1215.

Ubeda, G. (2018). *Song clips in class using technology and making it relevant.* [Blog, 16th December.] Accessed on 14th December 2024: https://geraldineubeda.wordpress.com/2018/12/16/song-clips-in-class-using-technology-and-making-it-relevant/

UCLA: Statistical Consulting Group (n.d.). *R Library Contrast Coding Systems for categorical variables*. Accessed on 25th January 2025:

https://stats.oarc.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/#SIMPLE

Unopeia. (2022). *I teach French to my key stage 2 class and I am terrible at it, to be honest.* [Comment on the online forum post, *Thoughts on the teaching of MFL in primary*]. Reddit. Accessed on 16th December 2024: https://www.reddit.com/r/TeachingUK/comments/xznhe4/comment/irn7nrm/

Unyk, A.M., Trehub, S.E., Trainor, L.J., & Schellenberg, E.G. (1992). Lullabies and simplicity: A cross-cultural perspective. *Psychology of Music*, 20, 15–28.

van Orden, K. (2006). Children's voices: singing and literacy in sixteenth-century France. *Early Music History*, *25*, 209–256. doi:10.1017/S0261127906000179

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

Vaissière, J., & Michaud, A. (2006). Prosodic constituents in French: A data-driven approach. In I. Fónagy, Y. Kawaguchi, & T. Moriguchi (Series Ed.), *Prosody and Syntax*, (pp. 47–64). John Benjamins.

Vihman, M.M., Davis, B.L., & DePaolis, R.A. (1995). Prosodic analysis of babbling and first words: A comparison of English and French. *Proceedings of the XIIIth International Congress of Phonetic Sciences – Stockholm*, Vol. 4, 14– 27.

Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Viviani, E., Ramscar, M., & Wonnacott, E. (2024). The Effects of Linear Order in Category Learning: Some Replications of Ramscar et al. (2010) and Their Implications for Replicating Training Studies. *Cognitive Science, 48*(5), e13445. https://doi.org/10.1111/cogs.13445

Wade, P., Marshall, H., & O'Donnell, S. (2009). *Primary modern foreign languages. Longitudinal survey of implementation of national entitlement to language learning at Key Stage, 2*. https://core.ac.uk/download/pdf/4160418.pdf Accessed on 11th December 2024.

Wagner, M., & Watson, D.G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*(7–9), 905–945. https://doi.org/10.1080/01690961003589492

Wakefield, J.R., Doughtie, E.B., & Yom, L. (1974). Identification of structural components of an unknown language. *Journal of Psycholinguistic Research*, 3, 262–269.

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.

Wang, Y., Bergeson, T.R., & Houston, D.M. (2018). Preference for infant-directed speech in infants with hearing aids: Effects of early auditory experience. *Journal of Speech, Language, and Hearing Research, 61*(9), 2431–2439.

Waterhouse, L. (2006). Multiple Intelligences, the Mozart Effect, and Emotional Intelligence: A Critical Review. *Educational Psychologist*, *41*(4), 207–225. https://doi.org/10.1207/s15326985ep4104_1

Weinstein, N., & Baldwin, D. (2024). Reification of infant-directed speech? Exploring assumptions shaping infant-directed speech research. *Culture & Psychology, 30*(1), 216–242. https://doi.org/10.1177/1354067X221147683

Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *J. Phon, 34*, 343–371.

Wermke, K. (2002). *Untersuchung der Melodieentwicklung im Säuglingsschrei von monozygoten Zwillingen in den ersten 5 Lebensmonaten. [Investigation of cry melody*

development of monozygotic twins within the first 5 months of life.]: Professorial dissertation (Habilitation), Humboldt-University of Berlin. http://edoc.hu-berlin.de

Wermke, K., and Friederici, A.D. (2004). Developmental changes of infant cries – The evolution of complex vocalizations. Behav. *Brain Sci. 27*, 474–475.

Wermke, K, & Mende, W. (1994). Ontogenetic development of infant cry- and non-cry vocalizations as early stages of speech abilities. In R. Aulanko & A.M. Korpijaakko-Huuhka (Eds.), *Proceedings of the Third Congress of the International Clinical Phonetics and Linguistics Association, 9–11 August, Helsinki* (pp.181–89). Helsinki: Publications of the Department of Phonetics 39, University of Helsinki.

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.

Wermke, K., Robb, M.P. & Schluter, P.J. (2021). Melody complexity of infants' cry and non-cry vocalisations increases across the first six months. *Scientific Reports, 11*(1), 4137. https://doi.org/10.1038/s41598-021-83564-8

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., Ma, L., Chang, H., Liu, Y., Hesse, V. & Shu, H. (2017). Fundamental frequency variation in crying of Mandarin and German neonates. *Journal of Voice, 31*(2), 255–e25.

Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P.J., Wermke, P., Robb, M., Keller, H., & Lamm, B. (2016). Fundamental frequency variation within neonatal crying: Does ambient language matter? *Speech, Language and Hearing, 19*(4), 211–217. https://doi.org/10.1080/2050571X.2016.1187903

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.

Wiese, R. (1996). *The Phonology of German*. Oxford: Clarendon Press.

Wiliam, D. (2018). [Social media post]. "Nothing. It's just that there is very little evidence that it works. Doing things that might work, while not doing things that are known to work is, in my view, unprofessional, like withholding effective treatments for illnesses in order to research others." Available at: https://twitter.com/dylanwiliam/status/1028967287372632069

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

Willis, J., & Willis, D. (1996). Challenge and Change in Language Education. Oxford: Heinemann.

Winkworth, A., Davis, P., Adams, R., & Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech, Language, and Hearing Research, 38*(1),124–144.

Woodrow, H. (1951). Time Perception. In S.S. Stevens (Ed.) *Handbook of Experimental Psychology*, pp.1224–1236, New York: Wiley.

Wonnacott, E., & Viviani, E. (no date). [Web page]. https://github.com/n400peanuts/languagelearninglab/tree/master/tools Accessed 20th February 2024.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Xu, Y., & Liu, F. (2012). Intrinsic coherence of prosodic and segmental aspects of speech. In O. Niebuhr & H. Pfitzinger (Eds.) *Understanding Prosody – The Role of Context, Function, and Communication.* Walter de Gruyter, pp. 1–26.

*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

Zangl, R., & Mills, D.L. (2007). Increased brain activity to infant-directed speech in 6-and 13-month-old infants. *Infancy, 11*(1), 31–62.

*Zhaku-Kondri, B. (2014). Using Song Lyrics in the Classroom: Assessing the Utility of Song Lyrics for the Acquisition of a Foreign Language. Proceedings of INTCESS14 – International Conference on Education and Social Sciences Proceedings. *California Folklore Quarterly, 1*(4), 1201–1210.

Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (Eds.) (2020). *Systematic Reviews in Educational Research Methodology, Perspectives and Application.* Wiesbaden: Springer.

Appendices

Appendix A: Systematic review appendices

Appendix A1: Systematic Review paper

The published systematic review paper is available to read here.

Appendix A2: 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 A3: 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?
	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?
Out	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

Appendix A4: Risk of Bias

Review of the quantitative, qualitative, and mixed methods designs component ratings.

1.1 Studies reporting quantitative data

There were 60 studies reporting quantitative findings that were assessed using the quantitative nonrandomised section of the MMAT. 51% of total possible responses in the whole section (153 out of 300) were 'Yes'. However, 'Yes' was the majority response in only two of the five sub-questions (selection bias and completeness of outcome data). 'Can't tell' was the main response for appropriateness of measure. Administration of intervention was split almost equally between 'Yes' (50%) and 'Can't tell' (48%). 'No' was the main response for whether confounders were accounted for in the design and analysis (73%).

Selection bias

It was unclear whether participants represented the target population in 52 of the studies, but these have been marked as 'Yes', providing there was a clear report of the sample itself, in recognition of the fact that educational research relies upon convenience sampling in many cases and that one or two school classes cannot adequately represent the whole population. This is a limitation of using the MMAT tool for educational research since very few included studies[16,17,24,47] report the inclusion and exclusion criteria used to select participants and therefore the tool loses some nuance in appraising risk of bias of studies, giving some the benefit of the doubt if 'Yes' is applied, but being too strict on an otherwise well-reported paper if 'Can't tell' is applied. Overall, 100% of included papers had low risk of selection bias.

Intervention and outcomes measures

'Can't tell' represents 83% (n = 50) of the responses for appropriateness of outcome measures, since many papers used a researcher-devised measure rather than a standardised or validated instrument. Three studies[13,23,27] used an inappropriate measure of linguistic outcomes (e.g., a Likert scale questionnaire to measure students' vocabulary acquisition). Seven studies used a standardised, reliable instrument such as the PPVT[1,39,54], TOPEL[9], TOLD:P-3[17], or validated reading measures[25,30].

Reporting of data

'Yes' represents 83% (n = 50) of responses regarding attrition rates, with most papers adequately reporting completeness of outcome data, reporting negligible attrition (no higher than 20%, which is the benchmark set for this review following Petticrew and Roberts, 2006) and giving reasons for dropouts, or else not reporting any attrition and having consistency throughout statistical reports about the number of participants so that it was

possible to ascertain that no attrition had occurred. Seven papers[5,10,13,34,44,53,60] failed to report attrition or numbers of participants consistently or accurately enough to answer this question. Three papers[4,27,41] had attrition rates above the 20% benchmark, with Alley (1988)[4] falling to 66% participation in some of the unit tests, and Ludke (2010)[41] falling to 75% participation with 16 missing data points, just below the 80% benchmark. Gorijan, Hayati and Barazandeh (2012)[27] report 60 participants in the abstract, and variously 56 or 28 per group at pretest, thus it is difficult to know whether there was attrition or simply poor reporting involved.

Potential confounders

For the question of whether confounders were accounted for in the studies' design and analysis, 73% (n = 44) of papers received a 'No', and a further 17% (n = 10) 'Can't tell', thus only 10% of papers adequately accounted for confounding factors and had low RoB in this section. Arguably it is impossible to account for all confounding factors in this kind of research because we cannot easily control for hearing music outside of school (e.g., on the radio or YouTube), especially when measuring English as an L2 because English music is popular around the world. However, the authors considered that measuring and clearly reporting participants' cognitive abilities, existing L2 knowledge, language background, parents' education and socio-economic status, age, gender, years of schooling, musical aptitude or years of music training would be some appropriate confounding factors to account for and measurable with baseline tests and background questionnaires. Only one paper[12] accounted for most of these factors, with a further five papers[1,14,17,30,41] accounting for key factors appropriate to their research design. Most of the 42 papers that received a 'No' for this question failed to report clearly[2,5,9,56] or gather[3,7,10,13,15,18,20,21,22,23,26,27,28,29,31,32,33,34,35,39,42,43,44,45,47,48,49,50,51,52,53,55,58,59,60] any baseline measures other than testing prior knowledge of the target words or reporting a vague notion of groups having similar levels of L2 knowledge[6,9,22,28].

Several other confounders were recorded. Three longitudinal studies (lasting one[10,13] or two[11] years) did not account or control for any confounding factors, making any findings limited as they could be attributed to participants getting older or any number of environmental factors. Two studies[6,55] encouraged participants to listen to the intervention music outside of class time, thus time on task is unaccounted for. Chou (2014)[19] used a mixture of songs and games and confounded the independent variables thus, with no control group, the findings cannot be attributed to songs. Coyle and Gómez Gracia's (2014)[20] participants took the same test six times, which may have confounded results with practice

effects. Overall, the RoB in this area is high and any findings need to be interpreted in the light of failure to control for multiple confounding factors.

Administration of intervention

'Yes' and 'Can't tell' are almost equally represented responses regarding the fidelity of administration of interventions. For 48% (n = 29) papers, there was insufficient detail provided to know whether an intervention had been administered as intended, and for 50% (n = 30) there was enough detail to give a 'Yes' response, indicating that no irregularities were reported. Only one paper[12] reports treatment fidelity explicitly. RoB in this area is therefore low to moderate.

1.2 Studies reporting qualitative data

In total, 12 studies gathered qualitative data[1,15,19,23,32,35,36,41,42,46, 47,52]. McCormack and Klopper (2016)[46] received a high RoB overall, despite low RoB assessments on their qualitative data collection methods, because they do not report clear research questions (the first of the screening questions). The same is true of Priester (2011)[52].

Regarding whether data collection methods were adequate to answer the research questions, two studies[23, 42] received a 'Can't tell' response. Luo (2019)[42] failed to describe the interview methods used in enough detail to answer whether interviews are an appropriate way to find out if songs help cultivate students' interest in learning English vocabulary. Diakou (2014)[23] used focus groups and self-report journal methods to ascertain whether participants developed their vocabulary acquisition. The other 10 studies had low RoB for this component.

On the three questions of whether findings are adequately derived from the data, results substantiated by data, and there is coherence between qualitative data sources, results are more mixed. Again, Luo (2019)[42] gives such a scant report overall that findings are neither derived nor interpreted from the data, and we cannot tell if there is any coherence between sources. Jarvis (2013)[32] does not give enough detail and receives 'Can't tell' on all three questions, and it is unclear how Chou (2014)[19] coded and interpreted data from observations, and findings are not adequately substantiated by data. All three studies received a high RoB rating overall. Likewise, Albaladejo et al. (2018)[1] does not report the interpretive framework used to guide findings or report in detail how results are substantiated with data, hence receiving a moderate RoB rating.

In summary, qualitative data was generally appropriate to answer research questions in these studies and collected using adequate methods in 10 out of 12 cases. The RoB increased in the interpretation of results and substantiation of findings with the data.

1.3 Studies reporting mixed methods

There were 12 studies reporting mixed methods, but four do not clarify the rationale for doing so[1,32,35,42]. Six studies[15,19,36,41,47,52] integrated the qualitative and quantitative data effectively to answer the research question, and five[1,23,32,42,35] did not. Only five studies[15,19,36,41,47,52] adequately interpret the combined outputs of qualitative and quantitative components, whilst six fail[1,23,42] or report too scantly to tell[15,19,32] and thus receive higher RoB ratings. Divergences and inconsistences between qualitative and quantitative results are adequately addressed in six cases[19,23,36,41,47,52] but three[31,42,35] fail to do so and two[1,15] fail to report in enough detail. Overall, six studies[1,15,23,32,35,42] receive high RoB ratings for their use of mixed methods, and four studies[36,41,47,52] receive low RoB for mixed methods but two of these[47,52] receive high RoB on the global weight of evidence rating because they do not state their research questions explicitly.

Appendix B: Appendices for Methodology

Appendix B1: Recruitment materials

B1a: Email to Headteachers

Dear [Headteacher],

I am a qualified French teacher and now PhD student from the Department of Education at the University of Oxford, investigating how effective songs are for teaching and learning foreign languages, under the supervision of Professor Victoria Murphy and Dr Hamish Chalmers.

I am writing to ask if you would kindly consider allowing me to conduct a short research study with your Year 3 students during school hours in September to October 2023. This study will form part of my doctoral thesis investigating the extent to which songs influence how children learn French word order. The study has been reviewed and received approval from Oxford's Central University Research Ethics Committee (CUREC; approval reference C1A-23-015).

The study will involve your Year 3 students in a specially devised 3-week French programme for beginner learners, with four groups of children learning French songs, French nursery rhymes, stories, or their usual French curriculum for 45 minutes per day for 3 weeks. I am an experienced teacher of French and will deliver engaging and interesting lessons in all four groups, which will hopefully enrich your students' experience of learning French. I will make the materials available to your school should you wish to use them again with your students after the research project.

Before beginning the teaching period, participating pupils will complete some one-on-one testing in two 30-minute sessions with me, at times during the school day that is agreed on a schedule with their class teacher. The assessments will involve some fun and age-appropriate cognitive reasoning, English/French vocabulary, and rhythm-copying tasks as well as a simple "listen and repeat-after-me" French speaking task. The French speaking task will be repeated immediately after the 3-week teaching period, and again 4–8 weeks after the teaching period, to allow for comparison of the four different methods of teaching and their effect on pupils' French learning over time.

I will provide you with a summary of the research findings once I have analysed the data, and can provide a copy of the research report once I have written up the paper for publication as part of my PhD thesis.

I have attached a detailed information sheet for your consideration. I would be happy to discuss the study further with you and answer any questions that you may have.

Thank you for your time and consideration of this research project. I look forward to hearing from you.

Best wishes,

Catherine Hamilton

B1b: Social media post

Does your primary school have three Year 3 classes who could help me complete my PhD research @OxfordDeptofEd where I'm investigating how effectively songs help children learn new languages? I would come in and teach French for 3 weeks in groups, and collect some data about how they are progressing at the beginning and end. More details ⬇️

With my supervisors @vmurphyox and @hwc001, I am seeking to collaborate with a 3-form entry primary school in the approx. Gloucestershire/Oxfordshire area who have beginner French learners in Year 3 that would like a fun introductory 3-week series of French lessons with me this June/July or September/October! ⬇️

The study will involve your Year 3 pupils in a specially devised 3-week French programme for beginner learners, with four groups of children learning French songs, French nursery rhymes, stories, or their usual French curriculum as a comparison, for 45 minutes per day for 3 weeks. ⬇️

I am an experienced teacher of French and will deliver engaging and interesting lessons in all four groups, which will hopefully enrich your students' early experience of learning French. I will make the materials available to your school should you wish to use them again with your students after the research project. ⬇️

Interested? Let's do some research together into an under-studied area of language education! Contact me on catherine.hamilton@education.ox.ac.uk if you would like more detail about the project, and I can send you the full information sheet for your consideration (no obligation to take part). Thank you for your time reading this. Please share if you can. Merci!

Appendix B2: Data collection timetable

This schedule shows the dates when I was in School 1 or School 2, and other work/holidays, over the data collection period.

W/c date		Mon	Tues	Weds	Thurs	Fri
4th Sept	AM				SchA distribute consent forms	
	PM					Meet SchB team
11th Sept	AM	SchA collect consent forms		SchA all day: collect baseline data 121		
	PM	SchB send/collect consent forms				
18th Sept	AM	SchA all day: collect baseline data 121				
	PM					
25th Sept	AM			SchB collect baseline data 121		
	PM		SchA intervention 22 mins x4 groups per afternoon			
2nd Oct	AM	SchB baselines	Cate in Oxford	SchB baselines		
	PM	SchA intervention		SchA intervention		
9th Oct	AM	SchB baselines		Cate in Oxford	Cate in Oxford	SchB baselines
	PM	SchA intervention				SchA intervention
16th Oct	AM	SchA all day: collect posttest data 121			Cate in Oxford	SchB baselines
	PM					
23rd Oct	AM	SchB baselines all day			Cate in Oxford	Cate in Oxford
	PM					
30th Oct	AM	Half term		Cate in Oxford	Cate in Oxford	Half term
	PM					
6th Nov	AM	SchB baselines all day			Cate in Oxford	SchB baselines
	PM					
13th Nov	AM				Cate in Oxford	
	PM	SchB intervention 22 mins x 4 groups				SchB intervention
20th Nov	AM	SchA delayed posttests (10 mins per participant)			Cate in Oxford	
	PM	SchB intervention				SchB intervention
27th Nov	AM	SchA delayed posttests		Cate in Oxford	Cate in Oxford	
	PM	SchB intervention				SchB intervention
4th Dec	AM	SchB all day: collect posttest data 121				
	PM					
11th Dec	AM	SchB all day: collect posttest data contingency				
	PM					
Jan 8th	AM	SchB delayed posttests (10 mins per participant)				
	PM					

Appendix B3: Information sheets

B3a: Information for Headteachers

Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools.

PARTICIPANT INFORMATION SHEET

Central University Research Ethics Committee Approval Reference: **C1A-23-015**

Introductory paragraph

You are being invited to take part in a research project. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

Why is this research being conducted?

Songs are popular resources for language teaching, but there is not much research into how songs specifically contribute to pupils' language learning outcomes like grammar, vocabulary, or pronunciation. This study investigates to what extent introducing French in the form of songs, nursery rhymes, or stories influences Year 3 children's ability to learn French grammar. We know that young children use prosody (the rhythm of language) to help learn English word order. This research study will contribute to our knowledge of how prosody works with/without melody to help children learn French word order at the start of their French learning journey.

Why have I been invited to take part?

You have been invited to take part because this study focuses on Year 3 pupils (aged 7–8 years old).

Do I have to take part?

No. It is up to you to decide whether to take part. If you decide to take part, you may withdraw yourself from the research, without giving a reason, and with no negative consequences, by advising us of this decision. The deadline by which you can withdraw any information you have contributed to the research is within two months of giving consent. The data collected up to the point of withdrawal will be destroyed immediately and will not be shared, used or published in any way.

What will happen to me if I take part in the research?

The school participating in this study will give permission for Catherine Hamilton, the principal researcher, to enter the school property to work with the Year 3 pupils whose parent/guardian has given consent for them to participate in the study.

The teacher participating in this study will:

- Disseminate the parent/guardian information sheet and consent form; collect the returned consent forms and store them in a safe place until Catherine comes to collect them
- Provide Catherine with participating pupils' year and month of birth, and sex
- Choose dates and times for Catherine to collect the initial data (baseline tests) in two 30-minute sessions with each participating pupil, and choose times when Catherine will teach 45 minutes of French with each group per day

- Before the initial data collection, introduce Catherine to the pupils and explain that she has come to do some activities one-to-one with the laptop and then teach some French in groups for a few weeks
- Provide a quiet place in the classroom (or in the hallway) where Catherine can work one-on-one with each pupil to collect the baseline data
- Provide a space for Catherine to teach the French activities with each of the four groups
- Provide Catherine with the standard French curriculum that would be covered during the three-week teaching period for her to teach to the 4th group

The children participating in this study will:

- Complete four short baseline assessments over two 30-minute sessions, spaced at least 30-minute apart to minimise any cognitive strain on the pupils). These assessments will involve pointing to a picture or word which Catherine reads aloud, listening to a rhythm sequence on a laptop computer and tapping along with the rhythm on the mousepad, and listening to a French sentence and then repeating it to the best of their ability.
- Participate in 45 minutes of age-appropriate and fun French lessons with Catherine each day for 3 weeks. There will be four groups of pupils focused on either singing, nursery rhymes, stories or the usual school French curriculum to allow for comparison between methods in the study. All materials will be made available after the study period so that pupils can enjoy the materials from the other groups too.
- Complete one more 'listen-and-repeat' French speaking task at the end of the 3-week teaching period, and a final 'listen-and-repeat' speaking task 4–8 weeks afterwards.
- For the French speaking task, with prior parent/guardian consent and child assent at the time of assessment, pupils will be audio recorded using a microphone and laptop to create an accurate record of each child's repeated sentences. No photographs will be taken.
- Participating pupils can ask to pause or stop the assessment at any time.

Timetable of the research project

Before teaching: baseline assessments	**3 weeks of French lessons**	**Immediate posttest at the end of 3 weeks**	**Delayed posttest 4–8 weeks after lessons**
Split into 2x 30-minute sessions spaced at least 30 minutes apart: 1x cognitive ability 1x English and French vocabulary knowledge 1x rhythmic ability 1x French 'listen-and-repeat' speaking task	45 minutes per day, per group, focused on: 1 = songs 2 = nursery rhymes 3 = stories 4 = usual French lesson plan	French 'listen-and-repeat' speaking task (30 minutes)	French 'listen-and-repeat' speaking task (30 minutes)

What are the possible disadvantages and risks in taking part?

- To reduce the risk of virus or Covid-19 transmission, Catherine will sanitise her hands, the desk area, and the laptop mouse pad between each assessment session.
- To minimise any disruption to children's instruction time and avoid them missing the same parts of their school day repeatedly, Catherine will agree the assessment and teaching schedule in advance with the class teacher.

- To reduce safety concerns, Catherine will conduct one-to-one assessments in an open area where another adult familiar to the pupils is nearby, for example in a quiet area of the classroom or in the hallway near the classroom. Catherine has completed safeguarding training, is a qualified teacher, and has an Enhanced DBS check.
- To reduce the risk of breach of confidentiality, all personal data will be anonymised and any participating schools, teachers, or pupils will not be identifiable in publication. All paper data collected for this study (such as consent forms) will be kept in a locked filing cabinet and turned into electronic files after collection. These files and other electronically gathered data will be encrypted and kept in the University of Oxford password-protected secure storage system. Only the researchers will have access to data gathered during this study.

Are there any benefits in taking part?

Participating children will benefit from time learning French with a specialist languages teacher, and have fun participating in interesting, age-appropriate activities as a whole group or in the one-to-one activities.

Participating schools will have the opportunity to work with a qualified French teacher and enhance their knowledge of applied linguistics research through planning and discussions. A summary of the research findings will be sent to participating schools, which they may be able to use to help plan their future MFL offering in school.

By agreeing to participate, you are helping to contribute to an emerging field of research that could have wider implications for teaching and learning languages in primary school.

Expenses and payments

There will be no payment for taking part in this research and any expenses will be absorbed by the researcher (for example, travel costs, production of materials).

What information will be collected and why is the collection of this information relevant for achieving the research objectives?

The information you provide during the study is the **research data**. Any research data from which you can be identified (e.g., your name, your school's name, your teachers' names, and your pupils' names, birth year and month, sex and audio recording) is known as **personal data**.

Personal data will be anonymised before publishing the research study such that you, and your school, teachers and pupils are not identifiable. Personal data will be stored in a locked filing cabinet at the Department of Education and digitised, then stored in the University of Oxford's password-protected online storage system (OneDrive) for 3 years after publication or public release of the work of the research.

Only the researchers will have access to the research data. Responsible members of the University of Oxford may be given access to data for monitoring and/or audit of the research.

Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will be written up in a doctoral thesis, academic publications, conference presentations, publicly available report and website publications. All data will be anonymised and no participants will be identifiable from these outputs.

The University of Oxford is committed to the dissemination of its research for the benefit of society and, in support of this commitment, has established an online archive of students' research outputs. A copy of Catherine Hamilton's doctoral thesis will therefore be deposited

both in print and online in the Oxford University Research Archive where it will be publicly available to facilitate its use in future research, thereby increasing its potential impact and use.

Data Protection

The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance web site at https://compliance.admin.ox.ac.uk/individual-rights.

Who is funding the research?

This research is part of Catherine's DPhil, which is a self-funded programme of study for which she receives no financial recompense.

Who has reviewed this research?

This research has received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: **C1A-23-015**).

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this research, please contact Catherine Hamilton on catherine.hamilton@education.ox.ac.uk or Professor Victoria Murphy on victoria.murphy@education.ox.ac.uk (University telephone number 01865 274042), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:

The Chair, Social Sciences & Humanities Interdivisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Boundary Brook House, Churchill Drive, Headington, Oxford OX3 7GB

Further Information and Contact Details

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact the primary researcher:

Catherine Hamilton
Department of Education
15 Norham Gardens, Oxford, OX2 6PY
University tel: 01865 274042
University email: catherine.hamilton@education.ox.ac.uk

B3b: Information for class teachers

Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools.

PARTICIPANT INFORMATION SHEET

Central University Research Ethics Committee Approval Reference: **C1A-23-015**

Introductory paragraph

You are being invited to take part in a research project. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

Why is this research being conducted?

Songs are popular resources for language teaching, but there is not much research into how songs specifically contribute to pupils' language learning outcomes like grammar, vocabulary, or pronunciation. This study investigates to what extent introducing French in the form of songs, nursery rhymes, or stories influences Year 3 children's ability to learn French grammar. We know that young children use prosody (the rhythm of language) to help learn English word order. This research study will contribute to our knowledge of how prosody works with/without melody to help children learn French word order at the start of their French learning journey.

Why have I been invited to take part?

You have been invited to take part because this study focuses on Year 3 pupils (aged 7–8 years old).

Do I have to take part?

No. It is up to you to decide whether to take part. If you decide to take part, you may withdraw yourself from the research, without giving a reason, and with no negative consequences, by advising us of this decision. The deadline by which you can withdraw any information you have contributed to the research is within two months of giving consent. The data collected up to the point of withdrawal will be destroyed immediately and will not be shared, used or published in any way.

What will happen to me if I take part in the research?

If you are happy to take part in the research you will be invited to:

- Disseminate the parent/guardian information sheet and consent form; collect the returned consent forms and store them in a safe place until Catherine Hamilton comes to collect them
- Provide Catherine with participating pupils' year and month of birth, and sex
- Choose dates and times for Catherine to collect the initial data (baseline tests) in two 30-minute sessions with each participating pupil, and choose times when Catherine will teach 45 minutes of French with each group per day for three consecutive weeks
- Before the initial data collection, introduce Catherine to the pupils and explain that she has come to do some activities one-to-one with the laptop and then teach some French in groups for a few weeks
- Provide a quiet place in the classroom (or in the hallway) where Catherine can work one-on-one with each pupil to collect the baseline data

- Provide a space for Catherine to teach the French activities with each of the four groups for 45 minutes per group, per day
- Provide Catherine with the standard French curriculum that would be covered during the three-week teaching period for her to teach to the 4th group

The children participating in this study will:

- Complete four short baseline assessments over two 30-minute sessions, spaced at least 30-minute apart to minimise any cognitive strain on the pupils). These assessments will involve pointing to a picture or word which Catherine reads aloud, listening to a rhythm sequence on a laptop computer and tapping along with the rhythm on the mousepad, and listening to a French sentence and then repeating it to the best of their ability.
- Participate in 45 minutes of age-appropriate and fun French lessons with Catherine each day for 3 weeks. There will be four groups of pupils focused on either singing, nursery rhymes, stories or the usual school French curriculum to allow for comparison between methods in the study. All materials will be made available after the study period so that pupils can enjoy the materials from the other groups too.
- Complete one more 'listen-and-repeat' French speaking task at the end of the 3-week teaching period, and a final 'listen-and-repeat' speaking task 4–8 weeks afterwards.
- For the French speaking task, with prior parent/guardian consent and child assent at the time of assessment, pupils will be audio recorded using a microphone and laptop to create an accurate record of each child's repeated sentences. No photographs will be taken.
- Participating pupils can ask to pause or stop the assessment at any time.

Timetable of the research project

Before teaching: baseline assessments	**3 weeks of French lessons**	**Immediate posttest at the end of 3 weeks**	**Delayed posttest 4–8 weeks after lessons**
Split into 2x 30-minute sessions spaced at least 30 minutes apart: 1x cognitive ability 1x English and French vocabulary knowledge 1x rhythmic ability 1x French 'listen-and-repeat' speaking task	45 minutes per day, per group, focused on: 1 = songs 2 = nursery rhymes 3 = stories 4 = usual French lesson plan	French 'listen-and-repeat' speaking task (30 minutes)	French 'listen-and-repeat' speaking task (30 minutes)

What are the possible disadvantages and risks in taking part?

- To minimise any disruption to your instruction time and avoid children missing the same parts of their school day repeatedly, Catherine will agree the assessment and teaching schedule with you in advance.
- To reduce the risk of virus or Covid-19 transmission, Catherine will sanitise her hands, the desk area, and the laptop mouse pad between each assessment session.
- To reduce safety concerns, Catherine will conduct one-to-one assessments in an open area where another adult familiar to the pupils is nearby, for example in a quiet area of the

classroom or in the hallway near the classroom. Catherine has completed safeguarding training, is a qualified teacher, and has an Enhanced DBS check.

- To reduce the risk of breach of confidentiality, all personal data will be anonymised and any participating schools, teachers, or pupils will not be identifiable in publication. All paper data collected for this study (such as consent forms) will be kept in a locked filing cabinet and turned into electronic files after collection. These digitised files and other electronically gathered data will be encrypted and kept in the University of Oxford password-protected secure storage system. Only the researchers will have access to data gathered during this study.

Are there any benefits in taking part?

Participating children will benefit from time learning French with a specialist languages teacher, and have fun participating in interesting, age-appropriate activities as a whole group or in the one-to-one activities.

You will have the opportunity to work with a qualified French teacher and enhance your knowledge of applied linguistics research through planning and discussions. A summary of the research findings will be sent to you, which may be helpful in planning your future MFL lessons.

By agreeing to participate, you are helping to contribute to an emerging field of research that could have wider implications for teaching and learning languages in primary school.

Expenses and payments

There will be no payment for taking part in this research and any expenses will be absorbed by the researcher (for example, travel costs, production of materials).

What information will be collected and why is the collection of this information relevant for achieving the research objectives?

The information you provide during the study is the **research data**. Any research data from which you can be identified (e.g., your name, your school's name, and your pupils' names, birth year and month, sex and audio recording) is known as **personal data**.

Personal data will be anonymised before publishing the research study such that you, your school, and your pupils are not identifiable. Personal data will be stored in a locked filing cabinet at the Department of Education and digitised, then stored in the University of Oxford's password-protected online storage system (OneDrive) for 3 years after publication or public release of the work of the research.

Only the researchers will have access to the research data. Responsible members of the University of Oxford may be given access to data for monitoring and/or audit of the research.

Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will be written up in a doctoral thesis, academic publications, conference presentations, publicly available report and website publications. All data will be anonymised and no participants will be identifiable from these outputs.

The University of Oxford is committed to the dissemination of its research for the benefit of society and, in support of this commitment, has established an online archive of students' research outputs. A copy of Catherine Hamilton's doctoral thesis will therefore be deposited both in print and online in the Oxford University Research Archive where it will be publicly

available to facilitate its use in future research, thereby increasing its potential impact and use.

Data Protection

The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance web site at https://compliance.admin.ox.ac.uk/individual-rights.

Who is funding the research?

This research is part of Catherine's DPhil, which is a self-funded programme of study for which she receives no financial recompense.

Who has reviewed this research?

This research has received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: **C1A-23-015)**.

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this research, please contact Catherine Hamilton on catherine.hamilton@education.ox.ac.uk or Professor Victoria Murphy on victoria.murphy@education.ox.ac.uk (University telephone number 01865 274042), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:

The Chair, Social Sciences & Humanities Interdivisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Boundary Brook House, Churchill Drive, Headington, Oxford OX3 7GB

Further Information and Contact Details

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact the primary researcher:

Catherine Hamilton
Department of Education
15 Norham Gardens, Oxford, OX2 6PY
University tel: 01865 274042
University email: catherine.hamilton@education.ox.ac.uk

B3c: Information for parent/guardians

Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools.

PARTICIPANT INFORMATION SHEET

Central University Research Ethics Committee Approval Reference: **C1A-23-015**

What are we trying to find out?

This research aims to learn more about how using songs, nursery rhymes or stories in French lessons helps children in the early stages of them learning French at school. Since songs are popular in language lessons, this research will help us understand what songs contribute to the process of learning a new language.

Why has my child been invited to take part?

Your child has been invited to take part because this study focuses on Year 3 pupils (aged 7–8 years old).

Does my child have to take part?

No. You can ask questions about the research before deciding whether to allow your child to take part. If you agree to participation, you may withdraw your child from the research, without giving a reason, and with no negative consequences, by advising us of this decision. The deadline by which you can withdraw any information your child has contributed to the research is within two months of giving consent. The data collected up to the point of withdrawal will be destroyed immediately and will not be shared, used or published in any way.

What will happen if my child takes part in the research?

Your child will participate in a specially created series of French lessons with the researcher, Catherine Hamilton, who is an experienced languages teacher. The lessons will take place every day for 3 weeks, for 45 minutes during the school day, with other children from Year 3. The lessons are designed to be a fun introduction to French and will involve some singing, nursery rhymes, stories and other French-related activities, depending on the group that your child is assigned to. All groups will benefit from having a carefully planned and engaging French lesson every day for 3 weeks.

Before the period of 3 weeks of French lessons, your child will complete a series of short assessments over two 30-minute sessions, in a quiet part of the classroom one-on-one with Catherine. These assessments will involve pointing to a picture or word which Catherine reads aloud, listening to a rhythm sequence on a laptop and tapping along with the rhythm on the mousepad, and listening to a French sentence and then repeating it to the best of their ability.

With your consent and your child's assent, this 'listen-and-repeat' French activity will be audio recorded on the laptop, to have an accurate record of your child's speaking for the purposes of analysis only. These recordings will not form any part of the final research output. The speaking activity and recording will be repeated at the end of the 3 weeks of lessons, and once more 4–8 weeks after the lessons, with your child's renewed assent on each occasion.

No photographs will be taken. Participating pupils can ask to pause or stop the assessment at any time.

What are the possible disadvantages and risks in taking part?

There are no serious risks to your child in taking part. You may wish to know that:

- The researcher, Catherine Hamilton, will minimise any interruption to your child's instruction or play times by agreeing the assessment and teaching schedule with the class teacher in advance.
- To reduce the risk of virus or Covid-19 transmission, Catherine will sanitise her hands, the desk area, and the laptop mouse pad between each assessment session.
- To reduce safety concerns, Catherine will conduct the one-on-one assessments in an open area where another adult familiar to your child is nearby, for example in a quiet area of the classroom or in the hallway near the classroom. Catherine has completed safeguarding training, is a qualified teacher, and has an Enhanced DBS check.
- Catherine will be introduced to your child in advance of the assessments and lessons, and all efforts will be made to ensure your child is comfortable and enjoying him/herself. If at any point your child feels uncomfortable during the assessment process, the assessment will be stopped immediately, without negative consequences.
- To reduce the risk of breach of confidentiality, all personal data will be anonymised and your child, their teachers, and their school will not be identifiable in publication.
- All data collected during this study will be anonymised and kept on the University of Oxford password-protected storage system. Only the researchers will have access to data gathered during this study. Paper consent forms will be kept in a locked filing cabinet, then digitised and shredded as soon as possible after collection.

Are there any benefits in taking part?

Your child will benefit from time spent learning French with a specialist languages teacher, and have fun participating in interesting, age-appropriate activities as a whole group. Most children benefit from having one-on-one time with a qualified teacher and the assessments are age appropriate, and designed to be fun. The whole process will contribute to your child's learning and will hopefully be enjoyable for them.

By agreeing to participate, you and your child are helping to contribute to an emerging field of research that could guide how we teach languages in primary school more widely.

What information will be collected and what happens to the data?

The information you and your child provide during the study is the **research data**. Any research data from which your child can be identified (e.g., their name, birth year and month, sex and audio recording, their school's name, their teachers' names,) is known as **personal data**.

Personal data will be anonymised before publishing the research study such that you, your child, their school, and teachers are not identifiable. Personal data will be stored in a locked filing cabinet at the Department of Education and digitised, then stored in the University of Oxford's password-protected online storage system (OneDrive) for 3 years after publication or public release of the work of the research. Only the researchers will have access to the research data. Responsible members of the University of Oxford may be given access to data for monitoring and/or audit of the research.

Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will be written up in a doctoral thesis, academic publications, conference presentations, publicly available report, and website publications. All data will be anonymised and no participants will be identifiable from these outputs.

The University of Oxford is committed to the dissemination of its research for the benefit of society and, in support of this commitment, has established an online archive of students' research outputs. A copy of Catherine Hamilton's doctoral thesis will therefore be deposited both in print and online in the Oxford University Research Archive where it will be publicly available to facilitate its use in future research, thereby increasing its potential impact and use.

Data Protection

The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance web site at https://compliance.admin.ox.ac.uk/individual-rights.

Who is funding the research?

This research is part of Catherine's DPhil in Education, which is a self-funded programme of study for which she receives no financial recompense.

Who has reviewed this research?

This research has received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: **C1A-23-015**).

Your child's assent

To help your child decide whether he/she would like to participate in this study, we ask that you discuss the study with him/her. You could use the following text, which will also be read to your child before beginning the one-on-one assessment sessions.

> I am a teacher and I want to know more about how children learn French. If you decide to help, I will ask you to do some little puzzles and point to words when I say them aloud, and tap along to a rhythm game that plays on the laptop. It is okay if you do not know the answers to the puzzles. I am just interested in hearing what you think and say, and having fun with the rhythm game.
>
> I will also ask you to listen to a sentence in French, and then try to repeat it as best you can. I would like to record your voice for this part, if that's okay with you? We will use this microphone and computer. The recording will help me later when I'm trying to remember exactly what you said.
>
> Then I'm going to teach you and your classmates French every day for a few weeks, doing lots of fun activities together. After that, we will do the French sentences game again where I say something, and you copy me. And then a few weeks later I'll come back to visit, and we will do the French sentences again if you are happy to join me.
>
> If at any time you don't want to do any more activities with me, you can just say you want to stop, and we will stop straightaway. No one will be upset with you if you want to stop. It will be okay if you want to stop at any time or have a little break and try again later. You can just let me know, okay?

What should I do next?

Please fill in the enclosed form and **return it to your child's class teacher** if you would like your child to take part in this study. Please remember that you may withdraw your child within two months of giving consent, without any penalty and without giving a reason, by notifying the researcher.

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this research, please contact Catherine Hamilton on catherine.hamilton@education.ox.ac.uk or Professor Victoria Murphy on victoria.murphy@education.ox.ac.uk (University telephone number 01865 274042), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:

The Chair, Social Sciences & Humanities Interdivisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Boundary Brook House, Churchill Drive, Headington, Oxford OX3 7GB

Further Information and Contact Details

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact the primary researcher:

Catherine Hamilton
Department of Education
15 Norham Gardens, Oxford, OX2 6PY
University tel: 01865 274042
University email: catherine.hamilton@education.ox.ac.uk

Appendix B4: Consent forms

B4a: Headteacher consent form

Consent to take part in
"Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools."

Central University Research Ethics Committee (CUREC) approval reference: **C1A-23-015**

The aim of this study is to investigate how songs contribute to pupils' learning of French word order in Year 3. By agreeing to participate, you are helping to contribute to an emerging field of research that could have wider implications for teaching and learning languages in primary school.

	Please initial each box if you agree with the statement
I confirm that I have read and understand the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
I understand that my participation is voluntary and that I am free to withdraw within two months of giving consent, without giving any reason, and without any adverse consequences or penalty.	☐
I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	☐
I understand that I will not be identifiable from any publications or presentations.	☐
I understand how to raise a concern or make a complaint.	☐
I agree to take part.	☐

_____	dd / mm / yyyy	_____
Name of participant	Date	Signature
Catherine Hamilton	dd / mm / yyyy	_____
Name of person taking consent	Date	Signature

B4b: Class teacher consent form

Consent to take part in
"Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools."

Central University Research Ethics Committee (CUREC) approval reference: **C1A-23-015**

The aim of this study is to investigate how songs contribute to pupils' learning of French word order in Year 3. By agreeing to participate, you are helping to contribute to an emerging field of research that could have wider implications for teaching and learning languages in primary school.

	Please initial each box if you agree with the statement
I confirm that I have read and understand the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
I understand that my participation is voluntary and that I am free to withdraw within two months of giving consent, without giving any reason, and without any adverse consequences or penalty.	☐
I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	☐
I understand that I will not be identifiable from any publications or presentations.	☐
I understand how to raise a concern or make a complaint.	☐
I agree to take part.	☐

_____	dd / mm / yyyy	_____
Name of participant	Date	Signature
Catherine Hamilton	dd / mm / yyyy	_____
Name of person taking consent	Date	Signature

B4c: Parent/guardian consent form

Consent to take part in
"Investigating the efficacy of whole-class singing activities for young learners of French in UK primary schools."

Central University Research Ethics Committee (CUREC) approval reference: **C1A-23-015**

The aim of this study is to investigate how songs contribute to pupils' learning of French word order in Year 3. By agreeing to participate, you are helping to contribute to an emerging field of research that could have wider implications for teaching and learning languages in primary school.

	Please initial each box if you agree with the statement
I confirm that I have read and understand the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
I understand that my child's participation is voluntary and that I am free to withdraw my child within two months of giving consent, without giving any reason, and without any adverse consequences or penalty.	☐
I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	☐
I understand that my child and I will not be identifiable from any publications or presentations.	☐
I consent to my child being audio recorded and my child assents to being audio recorded for the purpose of data analysis.	☐
I understand how to raise a concern or make a complaint.	☐
I agree to let my child take part.	☐

_____	dd / mm / yyyy	_____
Name of participant	Date	Signature
_____	dd / mm / yyyy	
Name of person taking consent	Date	Signature

Appendix B5: Language and music background questionnaire

Name	Code	Date
School	Class	Teacher
What languages can you speak? **Who else lives with you at home?** Mum/Caregiver 1 | Dad/Caregiver 2 | Brothers | Sisters | Anybody else? Tick: ……. …….. …… ……. ……..		

A: INTERPERSONAL INTERACTION	
A1. Which language does your Mum/Caregiver 1 speak at home?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A2. Which language does your Dad/Caregiver 2 speak at home?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A3. Which language do you use when you talk to your Mum/caregiver 1?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A4. Which language do you use when you talk to your Dad/caregiver 2?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
(If any other adults reported to live with the child, answer A5. If not, go to A6).	
A5. Which language do you use when you talk to any other adults who live at home?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A6. Which language do you use when you talk to your brothers and sisters?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A7. Overall, what language do you hear the most in your home?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	

A8. Are there any other children in this school who speak [insert L1]? Yes No (If yes, complete Question A8a. If No, go to Question A8b).	
A8a. Which language do you use in the playground with other children who speak [insert L1]?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
A8b. Do you know any other children outside your family who speak [insert L1]? Yes No (If yes, complete Question A8c. If No, go to Section B).	
A8c. Which language do you speak with these children?	**Score (circle)**
i. Always English	1
ii. English and L1 equally	2
iii. Mostly L1	3
Other	
Section A Total: Total score / (Number of times points awarded x3) =	
B: L1 MEDIA EXPOSURE	
B1. Do you ever watch television or films in [insert L1]? Yes No (If yes, complete Question B1a. If No, go to Question B2)	
B1a. Do you watch more English or [L1] television and films?	**Score (circle)**
i. More English	1
ii. Both equally	2
iii. More L1	3
B2. Do you ever hear the radio in [L1]? Yes No (If yes, complete Question B2a. If No, go to Question B3).	
B2a. Do you hear more English or [L1] radio?	**Score (circle)**
i. More English	1
ii. Both equally	2
iii. More L1	3
B3. Do you ever listen to music in [L1]? Yes No (If yes, complete Question B3a. If No, go to Section 4)	
B3a. Do you listen to more English or [L1] music?	**Score (circle)**
i. More English	1
ii. Both equally	2
iii. More L1	3
Section B Total: Score / 9 =	
SECTION C: MUSIC TRAINING	
C1. Are you learning any musical instruments? Yes No (If yes, continue. If No, finish)	
C2. How often do you have lessons on [instrument]?	**Score (circle)**
i. Less than once per week	1
ii. Once per week	2
iii. More than once per week	3

C3. How long have you been learning [instrument]?		**Score (circle)**
i.	Up to 1 year	1
ii.	1-2 years	2
iii.	More than 2 years	3
C4. How long do you practise [instrument] each week?		**Score (circle)**
i.	Less than 30 minutes	1
ii.	30 minutes	2
iii.	More than 30 minutes	3
	Section C Total: Score / 9 =	

Appendix B6: Pilot study report

B6.1 Pilot study research questions

The aim of this pilot study was to ascertain the feasibility of conducting a similar study on a larger scale with the eventual full cohort of participants. The research questions for the main study are detailed in section 3.3.1. This pilot study was guided by the following sub-questions:

1 How long does it take to administer the screening variables, and is this practical on a larger scale for a DPhil student?
2 How do the children respond to the screening/testing and French input materials, and are they appropriate for their age group?
3 Are the four input conditions (song, chant, story, control) comparable in terms of how they can be delivered, and how the children respond to them?
4 Is it possible to gather data using the Elicited Imitation Task materials/process that will permit the research questions to be addressed in the main study?

B6.2 Pilot participants

Sixteen children (mean age 103.88 months, *SD* = 7.35), 9 boys and 7 girls, participated in the pilot study from mid-May to mid-June 2023. Participants were recruited from the sampling frame constrained to the south-west region of England, within a 25km radius of my address to permit daily travel to/from the participating school. This sampling area contained 10 primary schools with three-form entry so that participants could be recruited from as few schools as possible to reduce environmental factors influencing the findings. These 10 schools' head teachers were invited to participate by email. The pilot school agreed to participate in the main study with their incoming Year 3 pupils from September 2023, and to participate in the pilot study with their current Year 3/4 composite class pupils.

Pupils who were in Year 4 at the time of the pilot had completed some Spanish lessons in Year 3, switching to French lessons in Year 4. The Year 3 pupils had completed some French lessons in Year 3. None of the pupils had received more than 3 hours' total school French lessons at the time of the pilot. Their existing French (or Spanish) lessons comprised of materials from the Primary Languages Network (PLN) courses that were organised and prepared by the MFL lead teacher and taught by their five regular class teachers, only one of whom is an MFL specialist who speaks Spanish and is learning French to keep ahead of the PLN materials.

Opt-in consent to participate in the pilot study was obtained from sixteen participants' parent/guardians. The class teachers selected participants from across their high, medium and low academic attainment bands to create a representative group of their whole cohort. All pupils spoke English at school. Four pupils mainly spoke languages other than English with their parents (three Polish and one Tamil), and one boy spoke Urdu and English equally at home. Participants were randomised into four treatment/control groups using www.random.org/lists random number generator, with the first four random numbers assigned to the song group, then the next four to the chant, the story, and the control groups respectively. Pilot participant characteristics are summarised in Table B6.1.

Table B6.1. Characteristics of the pilot participants

Group	**Male**	**Female**	**Mean age (months)**	**SD age (months)**	**EAL status**	**Total**
Song	3	1	101.00	5.48	2	4
Chant	1	3	100.75	7.50	2	4
Story	2	2	106.25	8.54	0	4
Control	3	1	107.50	7.85	1	4
Total	**9**	**7**	**103.88**	**7.35**	**5**	**16**

B6.3 Pilot screening variables

Standardised tests of cognitive ability (WASI FSIQ-2 Matrix and verbal reasoning), prior French knowledge (Échelle de Vocabulaire en Images; EVIP), and rhythm (children's rhythm synchronisation task; c-RST) were carried out to ascertain the comparability of the four groups. Due to school schedule constraints and to keep momentum going with the project between the screening and intervention stages after delays obtaining my DBS certificate, the screening data were analysed after the three-day intervention rather than before, as would be preferable to ascertain group comparability in the main study. Here it was decided that any group differences at baseline would not prevent the input and test materials being trialled for feasibility and the likelihood of the study gathering appropriate data to address the research questions could be ascertained regardless of any outcome measure group differences, which could only be ascribed to small group sizes or baseline differences with such small numbers. This pilot was a trial of the materials rather than a true experiment with inferential statistics sought to determine any effect of the intervention.

WASI FSIQ-2 measures of verbal and non-verbal reasoning

The WASI FSIQ-2 vocabulary test includes maximum 25 visual or written items for the 7–11 age group, presented in a booklet. For the 6–8 age group, the test begins with the first written item which is presented orally (i.e. "Can you tell me what a *shirt* is?") and the participant defines the word orally. I read the word from the spreadsheet that was created for ease of scoring and recorded the child"s response verbatim on the spreadsheet. If a child had not been able to define the first word (*shirt*), then the preceding visual items would have been presented, but this did not occur. If a child scored zero for five consecutive items, the test was stopped. After administering the test to all participants, I scored the responses again to double check the scores were consistent with the rubric, namely that responses with two accurate features receive two marks, one feature receives one mark, and inaccurate/no response/irrelevant responses receive zero. The vocabulary test took 10–15 minutes to complete, with some children asking questions about the words and telling stories that were prompted by the items. This helped build a rapport with the participants, but substantially increased the administration time of the baseline measures.

The Matrix reasoning test includes maximum 24 items. Participants select one image in the series of five options at the bottom of the page that they believe completes the matrix or series represented at the top of the page. The Matrix test was administered using the stimuli booklet, with children asked to either point at their choice of image to complete the matrix or to say which number image (1–5) they chose. I typed the chosen number directly into the Excel spreadsheet that was created for scoring the test. A formula was used to calculate when four consecutive answers (or four out of the previous five answers) scored zero. The rest of the row on the spreadsheet then turned red to indicate the test should be discontinued, which happened for eight of the participants (half of the group). The Matrix test took 5–10 minutes to complete per participant.

French vocabulary knowledge

The Échelle de vocabulaire en images (EVIP) French vocabulary test (Dunn et al., 1993) was presented to participants on paper. A page of the test booklet has four black and white images displayed, with participants asked to name one of the pictures orally (e.g. "Can you point to *le chien* for me?"). Participants were asked to point to the picture that best corresponded to the target word, or to give the number that corresponded to their choice of picture. The test took 5–10 minutes to administer to each participant. After eight consecutive errors, the ceiling level was reached and the test finished.

A spreadsheet was created to record the participants' responses, rather than using the paper EVIP test record sheets. To streamline the scoring process and reduce potential human error in scoring, a formula calculated when eight consecutive errors were made and then highlighted the rest of the spreadsheet row in red, so the researcher knew when to stop the test. This permitted the researcher to focus on establishing a good working relationship with each child and not be preoccupied with calculating the test score.

Children's Rhythm Synchronization Task (c-RST)

The c-RST was presented on a Windows Surface laptop, model i5, running the NeuroBehavioural Systems Presentation software package (version 23.1). The laptop had a mouse trackpad. Participants were shown how to click the trackpad during the practice phase of the task as a non-click would count as missing data. In future, it would be preferable to use a plug-in mouse with a left and right-click button, since one participant accidentally clicked out of the experiment by right-clicking on the trackpad too many times. This resulted in some data being lost even though he had almost completed the test trials.

It took 15–20 minutes to administer the c-RST. For the first six participants, no operational errors or interruptions occurred during test administration. For the 7th participant, the test was interrupted by lunch (due to me not having the correct lunch schedule for the year groups written down). It was then decided to not begin a new c-RST with any less than 25 minutes remaining before lunchtime for future participants. The 8th participant accidentally exited during the trials. The 9th participant experienced a frozen screen after the final test item and did not wish to complete the control items. The 14th participant had a frozen screen after the practice items, thus the test was restarted.

In general, the children found the test of low to medium interest, with a couple of exceptions who really enjoyed it. For the first ten participants, I went all the way through the practice, test and two control tasks, and found it necessary to gently motivate participants to continue. For the final six participants, the c-RST was broken into two phases: practice and test phase were completed, and then the two control tasks were completed after administering two other screening tests. By breaking up the c-RST in this way, the monotony of the test was reduced without interrupting the test trials and, it was felt, without reducing the test's reliability since all the tasks were independent of each other.

B6.4 Language background and music questionnaire

The language background and music questionnaire (see Appendix B5) was administered one-to-one with each participant, with answers recorded on a spreadsheet. It provided the opportunity to get to know the participant and for them to engage with me before embarking on the screening variable tests. They seemed to enjoy talking about their home languages and instrument lessons.

B6.5 Primary outcome measure: Elicited Imitation task

Pilot test materials

The EIT stimuli were recorded by the researcher, a native English speaker, using Apple AirPod headphones and the voice memos application on a MacBook (2017, 12-inch model) running OSMonterey. Using the audio insert feature, sound files were inserted into a PowerPoint presentation to administer the EI task. Each stimulus sound file was placed on its own slide, with the number of the stimulus (e.g. Practice 1, Sentence 20) written on the slide. Words contained in the audio recording were not presented as text on screen.

There were five practice and 22 test sentences, replicating earlier work with similar participants in Poland by Campfield and Murphy (2014). The practice sentences ranged in length from 2 to 6 syllables, and the test sentences from 4 to 9 syllables. It was anticipated that sentences longer than six syllables would present a considerable challenge to the new-to-French participants and permit differentiation between the groups by avoiding a low ceiling level being reached. Presenting longer EIT sentences permits learners to demonstrate their full ability range, whereas making the test too easy could erroneously restrict nuances in the highest performances (Phakiti, 2014). No particular language structures were targeted. Rather, learners' developing general structural knowledge of French was assessed to ascertain whether exposure to rhythmically salient input, with or without the addition of a melody, confers any advantage compared to prose input or typical MFL classroom language input in the 'business as usual' control condition.

EIT stimuli (see Table 4.5) have been constructed to avoid any advantage of group membership. Every stimulus contains lexical and structural items that appear across song, chant, and prose conditions, and from across the whole battery of input materials, thus not giving more advantage to input presented earlier or later in the three-week intervention period of the main study. Consequently, for this pilot study, some cue sentences will not have appeared in the input during the three-lesson (rather than the full three-week)

intervention. It was anticipated that pilot participants would find the sentences that they had previously encountered potentially easier to repeat than the novel sentences.

EIT Administration

The EI task was administered using the MacBook and a Snowball iCE Logitech microphone. Each sound file containing the five practice and 22 cue stimuli was played by pressing the "play" button on the numbered PowerPoint slide. Participants were told they could request each stimulus to be played one extra time.

Children were tested individually. At the pretest, the first three participants were tested in a quiet room with the class teacher present (marking work at a separate desk on the other side of the room where children had their back to her, to avoid distracting the participants) because my DBS certificate had not yet arrived and thus I could not work alone with the children. I sat at a round desk with the laptop on the left of the participants. The sound files were played through the MacBook speakers. The next day, the same constraints with the DBS arose and the pretests were administered at a round desk in the library, next to the class teacher's classroom. To avoid distraction from background noise, the five participants who took part on day two listened to the EIT using headphones. The remaining eight participants were administered the EIT after my DBS certificate had arrived, and thus listened via the MacBook speakers in a quiet room with only the researcher present. One participant who had used headphones at the pretest mentioned in the posttest that she would prefer to listen using the headphones as it was harder to hear without them. Whilst there seemed to be no difference to the other four participants who used headphones at pretest about the sound quality when using headphones or listening through the computer speakers, using headphones would be a better option to maintain similar sound quality for all participants.

Participants were instructed to 'say what they hear' and repeat the French as best as they could. They were told they were not expected to say what any of the French words mean, or provide a translation, just to try and imitate the cue. This was difficult for some of them to grasp and hence the five practice sentences were useful to train participants in the procedure. They could take as long as they wished to respond, with responses (or refusals) generally given within a few seconds of the stimuli being heard. The participants were told that their responses would be recorded so that the researcher could compare the responses from before the lessons and after the lessons, to see if there were any changes. Participants all assented to being recorded. However, two children were reluctant to do the task: one

stopped after the first sentence in the pretest saying that it was too difficult, and she did not want to continue. She completed the posttest. A second participant, keen on learning French outside school using the app Duolingo, found that he could not repeat the sentences because he was consciously trying to process their meaning, and felt unable to repeat stimuli that he did not explicitly understand. He said he did not want to get it wrong and could not proceed past the 4th cue sentence in the pretest. He also completed the posttest but still found it difficult to repeat sentences he did not explicitly understand because he did not like getting his French wrong. These two participants provide an insight into how extremely low-performing or high-performing participants may respond to the task, with both extremes resulting in refusal to perform. Most participants were willing to repeat the cue sentences without questioning what they meant, even if they were slightly baffled by the request to say things they did not understand the meaning of.

B6.6 Design of input materials and lesson plans

To test the planned input materials and lesson delivery, three lessons from each condition were prepared in full. Each group received 1.5 hours (30 minutes each day for three days) of French input. The song group were introduced to two songs per day, thus six songs. The chant group were introduced to the same six songs as the song group, without the melody. The story group were introduced to the first three instalments of the story. The control group received three lessons about animals using the PLN content that was recreated in a PowerPoint rather than presented directly from the PLN videos and materials online, as guided by the school's usual MFL practice.

Whilst it was impossible to match the vocabulary exactly across the song/chant and prose conditions when only three parts of the story were being introduced in the pilot, six songs were chosen that introduced two key characters in the story (the *lapin* and the *loup*) and therefore some vocabulary and structures were similar across conditions. Care was taken to choose songs that represented the difficulty spectrum of the whole repertoire of 24 songs: two were more repetitive and therefore perhaps easier to learn as they contained fewer new words and structures (*Savez-vous plantez les choux* and *Meunier, tu dors*); two were less repetitive but still reused similar structures (*Mon petit lapin* and *Dansons la capucine*); one contained no repetition and was arguably the most difficult to learn (*Promenons-nous dans les bois*); and one was familiar in English (*Incy Wincy Spider*) but new in French, presenting a challenge as the tune was not the same as in English and the lyrics were quite complex, even if the narrative was familiar. As anticipated, the children in the song and chant

conditions commented on the ease of learning the repetitive songs/rhymes compared to the less repetitive ones or the one that should have been familiar but was different to what they expected when hearing the title.

The input materials were presented on a Macbook laptop via PowerPoint presentations. Each song or chant appeared in written form on its own slide. Illustrations were used to illuminate the themes (e.g. drawings of rabbits to accompany the *Mon petit lapin* song/chant). The song and chant conditions were identical except for the sound files which contained me either singing or chanting the input. I recorded these using the Snowball Ice microphone and inserted them as sound files into the PowerPoints. The recordings were all made on the same afternoon to ensure a similar quality of voice across conditions, and to allow me to internalise the same rhythm for both song and chant. Only an impressionistic view of 'the same rhythm' was used, rather than a quantitative approach to maintaining the same rhythm across conditions. It was felt that this was sufficient to capture what teachers would naturally do when singing or chanting, and that artificially altering the rhythm would make the input less naturalistic and may confound the experiment. Care was taken not to speed up or slow down in a way that was noticeably different between these two similar conditions.

The story condition was likewise presented on PowerPoint, with three slides representing the three instalments of the story. The same cartoons were used to illustrate the story as the song/chant slides with additional illustrations where the content differed. Each line of the story appeared when the Enter key was pressed, with the sound file playing automatically for each line.

The 'business as usual' control group materials contained audio files presenting the French animal vocabulary (e.g., *un chat*) that appeared in written form with accompanying illustration on screen (using the same illustrations as the other groups). Lessons two and three built up to a short phrase with a variable slot filled with animal vocabulary items (e.g., *C'est un chat*, and *Mon animal préféré est un chat*). Each slide in the PowerPoint introduced an animal, or the target phrase with each animal in turn during lessons two and three. A short game that reused the vocabulary (e.g., a 'who is missing' game, where children produced the word for the missing animal picture in lesson one, a 'create a fantastical composite animal' like *un chien-lapin* in lesson two, or a 'guess my favourite animal' game in lesson three) was played once the phrases had been introduced and practised. Each lesson finished with the game, which recapped that day's vocabulary and structures.

Presentation

The premise of each song or chant was first briefly explained in English, e.g. "This is a song about a rabbit who is feeling sad, and he's told that if he goes jump jump jump he can have some of the herb thyme as a treat." The song or chant was then played on the sound file three times, as Campfield and Murphy (2014) had done, but this felt like too many times to play it before the children had a turn at repeating it. Therefore, on the second lesson, the input was played through twice, then the children were taught to repeat it in sections, line by line building up from the last word and 'back chaining' to maintain speed and fluency until they could repeat the whole line. The audio file was then played in full again and they tried to sing or chant along. It felt unnatural and too performatively demanding to present this final stage as an audio file only while the children sang or chanted, and so I sang or chanted along with them to be more encouraging and inclusive.

In the prose group, the story was briefly explained in English, e.g. 'There is a rabbit who feels sad and his friends the owl and the cuckoo try to cheer him up by suggesting going for a walk. The rabbit is scared he will meet the wolf but the owl says they won't meet the wolf and it will be okay." Then the French audio was played twice line by line, with children repeating the lines and back-chaining from the end of the lines to build up their speed and fluency. Once the whole instalment had been repeated twice through line by line, it was all repeated from the beginning together with the audio file and I speaking in unison with the children.

In the control group, the premise of the mini unit was presented as learning about animals in French and how to say what animals we prefer. The pronunciation of words was broken down, following the PLN material guidance, and pertinent phonics were taught explicitly (for example that *ch* in French is pronounced *sh*). One L1 Polish child pointed out that this phoneme-grapheme link was the same in Polish. The audio files on each slide were played twice, children asked to repeat the vocabulary, and then everyone repeated the phrases in unison and played the games.

B6.7 Observations during pilot lessons

This section addresses pilot research question (3) **Are the four input conditions (song, chant, story, control) comparable in terms of how they can be delivered, and how the children respond to them?**

Participants' attitudes were almost unanimously positive about learning French in all groups. The song and prose conditions required little explanation about why they were being asked to do some songs or a story in French: the children accepted these as routine class activities. The song group were happy to participate and enthusiastic every lesson. They were a mixed ability group and worked well together. One boy could not sing in tune, but this did not prevent him joining in enthusiastically each day. There were two boys with EAL in the group, one Polish speaker (who heard mostly Polish at home) and one Urdu speaker (who reported hearing English/Urdu equally at home). They both began repeating the songs as soon as they heard them, without hesitation. The Polish boy was absent on the third lesson and missed the posttest. The group said they found the song with the most repetition (*Savez-vous planter les choux?*) the most enjoyable and they enjoyed the tune, and they found the most linguistically complex, least predictably melodic song the hardest (*Promenons-nous dans les bois*). They all enjoyed the *Mon petit lapin* song from the first lesson best and remembered the songs really well on the third lesson recap. Nobody seemed fazed by the lyrics being presented as text on screen. They used the written lyrics to find their place as they sang along. On the second lesson, after practising the song twice through, I turned the screen away and said to try singing the song without reading the words. There was noticeably more hesitation as the children listened more intently to find their place. Having the written words on screen appeared to help with fluency, even if they were unlikely to be reading the words with comprehension after such a short exposure.

The chant group were happy to participate on the first lesson. However, the chants were apparently less appealing than the songs as the chant group noticed how many words were on screen and it felt more challenging to engage them in chanting so many words. The presence of the tune in the song condition seemed to distract from the challenge of the words themselves, perhaps because children could focus on the tune and hum along if they did not immediately grasp the words. In the chant condition, there was a sense that it felt difficult and less fun, and there was a pull towards using actions to accompany the chants (indeed these songs or chants would normally include actions). However, the plan was to avoid introducing a new variable (actions) as this could confound the results. Instead, I encouraged participants to tap along with the rhythm on the table if they wanted to, which three out of four did.

After the first chant lesson content was introduced, one child expressed his dislike for French, which contrasted with his enthusiasm about joining in at the start. This dramatic turn in attitude affected the group's enthusiasm levels generally. I decided to try and motivate the

children by showing them what rhymes would be learned on the next two lessons, and explaining that the purpose of the lessons was to test how well the materials worked with children their age. It was felt that the outcome of the lessons needed to be more explicit than in the song or prose groups, where the aim was implicit in learning a song or story. In the second and third chanting lessons, there was slightly more positivity from the more disengaged boy, but he continued to say he found it difficult. *Savez-vous planter les choux?* contained more repetition and was easier for the chant group to grasp. On the third lesson, I tested doing the actions to the songs for *Meunier, tu dors* and dancing in a round for *Dansons la capucine*. The two girls joined in for the dance (a third girl was absent and caught up the lesson on another day). However, using actions made the lesson feel more light hearted and fun, and they enjoyed themselves more. The boy joined in with the arm rolling actions for *Meunier, tu dors*. He repeated that he found French hard and couldn't remember it all, but he did produce a lot of the chants and seemed to remember them well. One girl who caught up after her absence enjoyed the accompanying actions and, when returning to class, she showed her teacher how to do the chant (with my support in chanting). Overall, it seemed that the chants needed to be accompanied with actions to make them as motivating as the songs. However, whilst as a teacher this felt like a natural thing to do (boost motivation by adding the actions where perceived necessary in the moment), as a researcher it could introduce bias to include actions and not to isolate the song or chant audio-only input as the variable. Yet if the chant condition is less motivating than the songs, that also introduces bias to the experiment as motivation levels would not be comparable. The way the chants would be received by a different group of children would also change how, as a teacher, I felt it was better to introduce them.

The story group worked well together and all talked about how the new material linked with their prior knowledge. They asked questions about links between Spanish, Italian and French and seemed to be keen linguists. They engaged well with repeating the story and enjoyed the challenge of saying lots of new French words. One girl had done some Spanish previously, and pronounced the written word *dos* ([*doe*] for *back* in French) as the Spanish number two [*doss*]. We therefore talked about how phoneme-grapheme links can be different across languages and how French has silent letters. They were all keen to know what happened next in the story, and I gave a synopsis of how the rabbit's journey turns out. They said it was a funny story and seemed to find the narrative engaging. One boy was absent on the third lesson, so he caught up afterwards on his own and was just as engaged as when part of the group.

Children in the control group were more boisterous that in the other three groups, and harder to settle to task. In the first lesson they called out answers and were enthusiastic but did not listen as well as in the other groups. The Polish EAL student could roll his Rs in *serpent* and *araignée* and he taught the others how to do it, which caused some hilarity, but they were motivated to try and were having fun. The group was easier to settle in the second lesson, possibly because it was the first lesson of the day and not just after break-time as in lesson one. They were less chatty and more engaged and focused on the task, and apparently motivated by having remembered many of the animals from the day before. On the third lesson they were very unsettled and found it hard to listen to each other in the "guess my favourite animal" game (hence they kept repeating guesses that other people had said). Nonetheless, they remembered a lot of the words from previous lessons and were enthusiastic about showing off their knowledge.

Overall, all four groups participated willingly and generally very enthusiastically with the lessons. After the posttest, they were asked what they thought of the materials and of learning French. The boy who was disengaged in the chant group said he would give French 8.2 out of 10 because he found the chants "a bit tricky". The other fifteen present at posttest said they enjoyed the lessons and would like to do more French in the future.

B6.8 Pilot study results

This section reports the results of analyses conducted to ascertain group comparability, testing the feasibility of using the same screening measures and language background questionnaire in the main study. Then follow the results of the primary outcome measure, the Elicited Imitation Task, at pre- and posttest, with a view to deciding whether this data collection method garners data that permit the research questions to be addressed.

Screening variables

Language background and music questionnaire

The questionnaire worked well for gathering data about the level of L1 used outside of school, and provided an opportunity to get acquainted with the participants before beginning the testing.

WASI FSIQ-2 measures of non-verbal and verbal reasoning

The WASI Matrix non-verbal reasoning test was scored out of 28. Participants' raw scores ranged from 4 to 27, with an overall mean of 16.75 (SD=7.57). To permit comparison of scores from children of different ages, the WASI Matrix raw scores were converted to

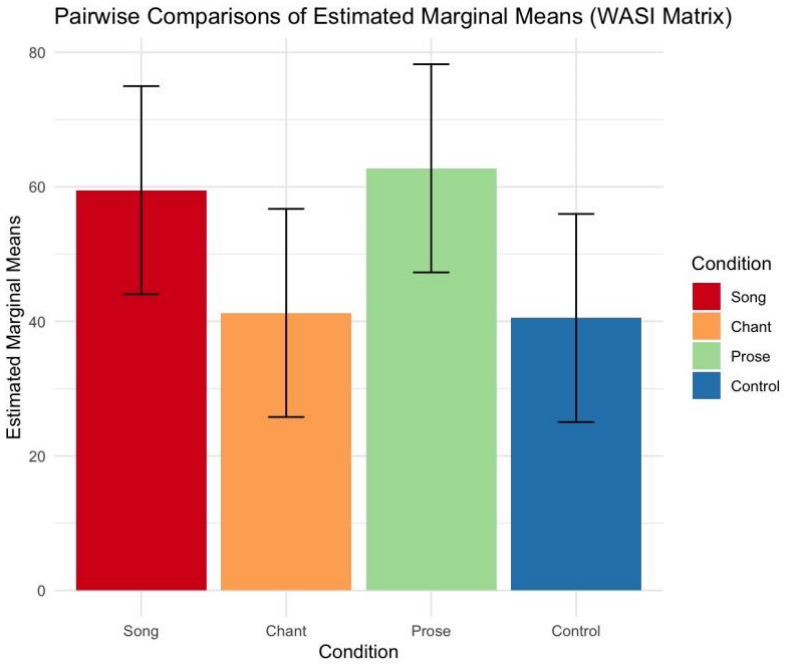
standardised t-scores. These ranged from 28 to 70, with an overall mean of 51 (SD=13.27). Group t-scores are summarised in Table B6.2.

Table B6.2. Summary of WASI Matrix pilot group t-scores

Group	*n*	Mean	SD
Song	4	59.50	2.89
Chant	4	41.25	12.42
Prose	4	62.75	5.62
Control	4	40.50	11.47

To ascertain whether differences in the t-scores between groups reached a level of statistical significance, a linear model comparison with Bonferroni correction for six tests was used. There were no significant pairwise differences between groups at the adjusted alpha level of .008. Figure B6.1 visualises the pairwise comparisons of estimated marginal means in group cognitive scores, as measured by the WASI matrix non-verbal reasoning test.

Figure B6.1. Plot of estimated marginal means in pilot group WASI matrix scores



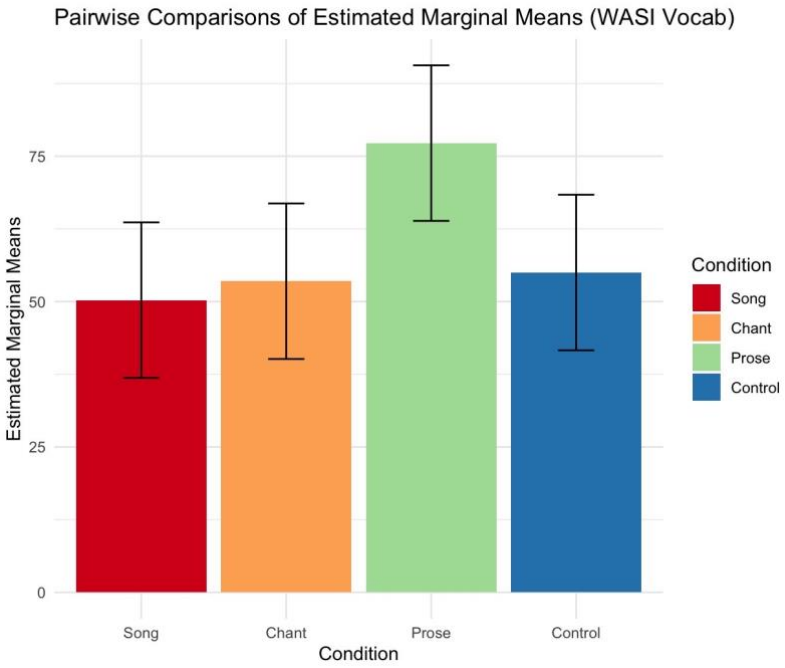
The WASI vocabulary verbal reasoning test was scored out of 60. Participants' scores ranged from 23 to 55, with an overall mean of 36.44 (*SD* = 10.22). The raw scores were converted to standardised *t*-scores to permit comparison across the age range of participants. The mean overall *t*-score was 59 (*SD* = 13.05). Table B6.3 shows the scores for each group.

Table B6.3. Summary of WASI Vocabulary pilot group t-scores

Group	*n*	M	SD
Song	4	50.25	7.46
Chant	4	53.50	5.80
Prose	4	77.25	4.86
Control	4	55.00	11.43

A linear model comparison (with Bonferroni adjusted alpha of .008) confirmed that the difference in WASI vocabulary scores between the Song and Prose (p = .002), and the Chant and Prose (p = .006) groups was statistically significant. Figure B6.2 illustrates the pairwise comparisons of estimated marginal means for the English vocabulary scores as measured by the WASI vocabulary test at baseline. Given that there were only four participants in each group for the pilot study, this comparison only serves as a test of the procedure for the main experiment rather than a reliable statistical test of the pilot baseline group comparability.

Figure B6.2. Pairwise comparisons of estimated marginal means for pilot WASI vocabulary



French vocabulary knowledge

Since none of the participants was a native French speaker, and the EVIP receptive French vocabulary knowledge test is intended for use with native speakers, it was scored in an alternative manner to the guidance in the manual. Rather than establishing a base vocabulary level after eight consecutive correct responses, each participant began the test at the first item (indicated as appropriate for children from age two years in the guidance). The data

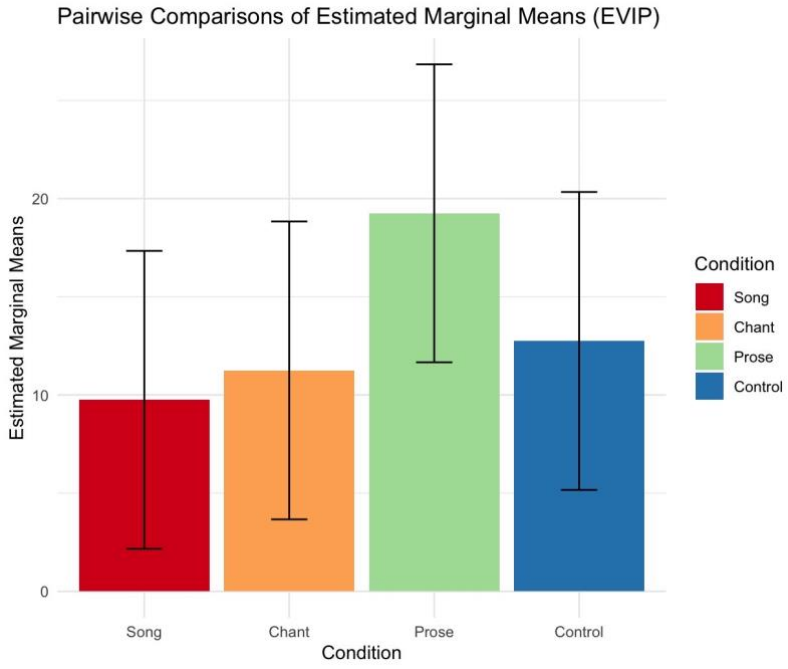
reported are the raw scores (number of correct items) rather than the normalised scores for age group since none of the participants reached the 1st percentile for their age, as might be expected for a group of non-native beginner L2 learners. Table B6.4 summarises the group scores. Participants scored between 6 and 22 correct items, with an overall mean of 13.35 (*SD* = 5.45).

Table B6.4. Summary of EVIP French vocabulary pilot group scores

Group	*n*	M	SD
Song	4	9.75	4.32
Chant	4	11.25	6.85
Prose	4	19.25	5.50
Control	4	12.75	4.50

A linear model comparison of the groups' French vocabulary scores was carried out, with Bonferroni correction for multiple pairwise comparisons. There were no statistically significant group differences at the adjusted alpha level of .008. Figure B6.3 illustrates the estimated marginal means on the baseline French vocabulary measure.

Figure B6.3. Pairwise comparisons of estimated marginal means for pilot French vocabulary



Children's Rhythm Synchronization Task (c-RST)

The c-RST produces two scores per participant: (1) their percentage correct score indicates how many taps they made within the 'scoring window' of half an interval before and after the stimulus, and (2) their inter-tap interval synchrony (ITI) indicates how accurately they reproduced the temporal structure of each rhythm. The c-RST Python script automatically calculates these scores and produces an output file with the data per participant. In this pilot, two participants (one in the Song, and one in the Chant group) had no corresponding output after running the Python analysis code, despite having logs of their responses in the files. According to the task creators' scoring notes, taps that do not fall within the scoring window are considered as missing and suppressed in the output.

Table B6.5. Summary of c-RST scores by pilot group

		% correct		**ITI**	
Group	***n***	**Mean**	**SD**	**Mean**	**SD**
Song	3	0.68	0.09	0.55	0.23
Chant	3	0.70	0.08	0.58	0.17
Prose	4	0.73	0.08	0.48	0.18
Control	4	0.77	0.11	0.40	0.21

Two linear model comparisons, one of the percentage of correct taps and one of the ITI synchrony, found no significant differences between groups on these two measures of rhythm synchronicity. Figures B6.4 and B6.5 illustrate the estimated marginal means of each measure.

Figure B6.4. Pairwise comparisons of estimated marginal means for pilot % correct rhythm taps

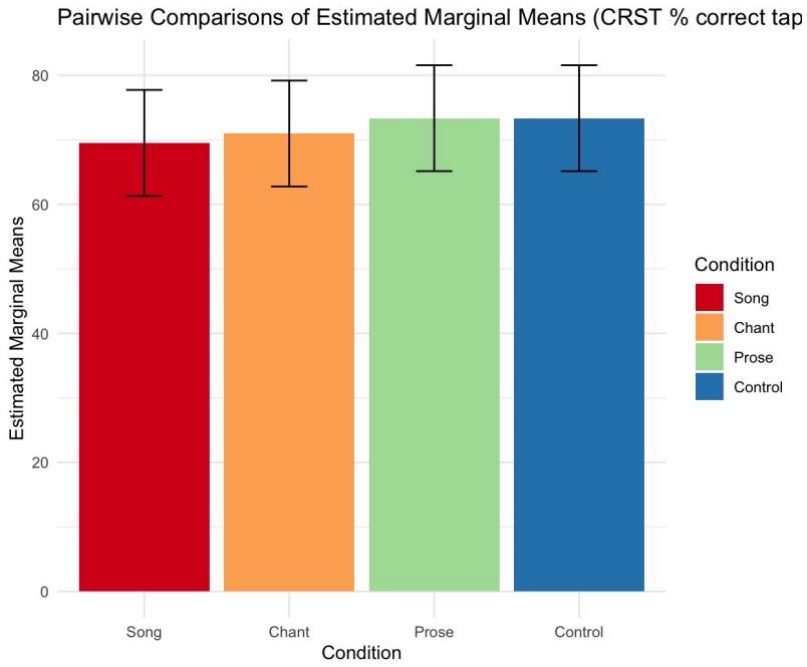
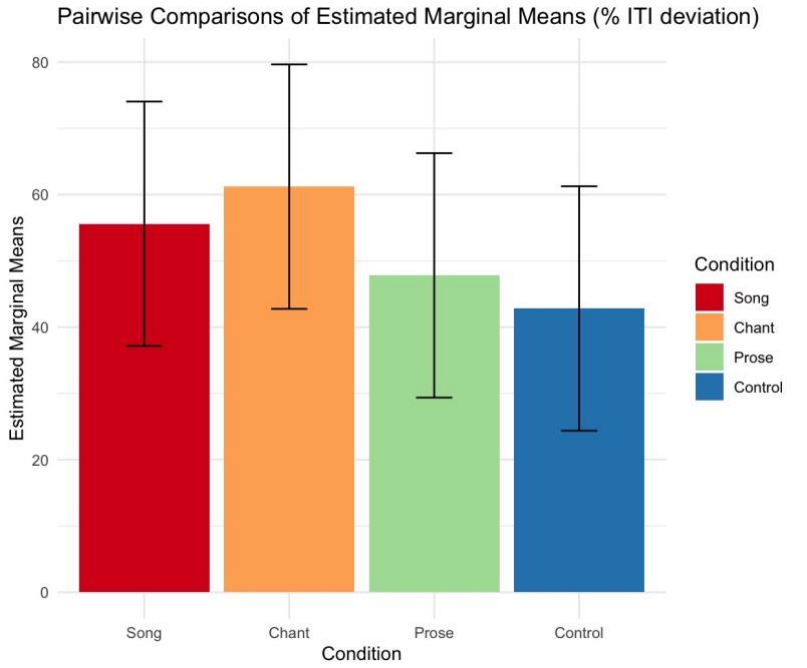


Figure B6.5. Pairwise comparisons of estimated marginal means for pilot ITI synchrony



Discussion of screening variables findings and suitability

The piloting of the screening variables was driven by two research questions, which are addressed in the following section.

1. **How long does it take to administer the screening variables, and is this practical on a larger scale for a DPhil student?**

The four screening measures (WASI matrix, verbal reasoning, EVIP and c-RST) produced data that permitted group comparisons to be made. Whilst no reliable inferences can be drawn from such a small sample size for this pilot study, the process of administering the measures demonstrated that they were suitable for the age group. However, the time it took to administer the screening variables and pretest was almost one hour per child, and on a larger scale this would not be practical or desirable. The WASI matrix and EVIP took five minutes each on average per participant. The c-RST was slightly longer, taking 15 minutes per participant. The WASI verbal reasoning test substantially increased the time required to administrate the screening variables as it took ten to fifteen minutes per participant and could only be administered on a one-to-one basis. This would be impractical on a larger scale, and excessively time-consuming for a measure that is not the focal point of the study. Even though the measure helped to build a rapport with the participants, this could be achieved through a brief and friendly discussion on the way from the classroom to the room where the testing takes place, and also during the administration of the language background questionnaire.

In discussion with colleagues about the limitations of the WASI verbal reasoning measure, an alternative test of participants' English vocabulary knowledge was found that could be administered as an online task to a whole class, taking no more than 30 minutes per group including the time to set up the task on school tablets and laptops. This Picture Vocabulary Size Test (PVST) is a software program with a receptive vocabulary test containing picture choices with oral cues developed by Anthony and Nation (2021). The 20,000-word level version is suitable for use with literate or pre-literate native or non-native speakers of English of any age. It was therefore deemed appropriate for the 7–8-year-old participants of this study. I confirmed with both participating schools that laptops or tablets would be available to administer the task in whole-class or small groups. I ran the PVST myself and trialled it with my 7-year-old son who was keen to participate in the research preparations. Having established that it worked on this small scale, and hearing positive accounts of the PVST from colleagues who were using it in their ongoing research projects, I decided to use it instead of the WASI verbal reasoning measure in the main experiment.

With this change, the time required for the screening variables in the main experiment would be as follows:

- With individual participants: 20 minutes (EVIP and c-RST)

- With groups/classes: 25 minutes (PVST and WASI matrix)

The process could be further expedited by administering the c-RST with headphones on a laptop to one child, whilst administering the EVIP and WASI matrix to a second child.

2. **How do the children respond to the screening/testing and French input materials, and are they appropriate for their age group?**

The children responded positively to the screening measures. The WASI matrix was received positively by the children, who enjoyed the challenge. Several likened it to the 11+ grammar school test that their older siblings had done in Year 6 (the region has an academically selective grammar school system for secondary school selection). Most children particularly enjoyed the Rhythm Synchronisation Task (c-RST) on the laptop, which they joyfully described to their classmates as a fun computer game. Two children found the c-RST rather dull and found it difficult to concentrate for the whole 15 minutes. The c-RST has a natural split between the training/test phase and the control tasks. I therefore, for those children who expressed boredom during the test phase, split the c-RST into two blocks and administered either the WASI matrix or EVIP in between the test phase and control tasks.

Some children expressed a lack of confidence about their French knowledge during the EVIP. I explained that I did not want to assume that, just because they had received no French lessons yet, they couldn't use their existing knowledge of English and other languages to work out the meaning of French words. I was able to reassure them that the EVIP was not a test of their existing French knowledge learned in school, but a way of seeing how well they could work out what French words might mean if they had not heard them before. They enjoyed the task much more once they knew it was not trying to test their existing knowledge, which as beginners would have been an unfair proposition. Two children expressed that they were 'just guessing' the answers, which I explained was a valid approach and used their language detective skills.

Elicited Imitation Task results

This section addresses the fourth and final pilot study research question:

5 **Is it possible to gather data using the Elicited Imitation Task materials/process that will permit the research questions to be addressed in the main study?**

The EIT gathered the participants' responses to the stimuli as audio data files. These were then transcribed and coded using the scoring scale of 0 (omission) to 5 (exact imitation; see 4.5.1 for full details about transcription and scoring).

The first test of whether the data gathered through the EIT process would permit the research questions to be addressed was to decide whether the proposed data analysis plan would result in a well-fitting model. The pilot data were modelled using a cumulative link mixed model (CLMM; see 4.5.2 for analysis plan). To determine whether the inclusion of random effects significantly improved the model fit, a likelihood ratio test was conducted comparing a fixed effects model, and a mixed effects model with both fixed and random effects. The ANOVA comparison of the two models is summarised in Table B6.6. The ANOVA indicated that the mixed effects model provided a significantly improved fit to the data ($p < .001$) compared to the fixed effects model.

Table B6.6. ANOVA comparison of fixed and random effects models

Model	**No. parameters**	**Akaike Information Criterion (AIC)**	**Log-likelihood**	**Likelihood ratio**	**df**	***p*-value**
Fixed effects	13	2265.4	-1119.71			
Mixed effects	52	1986.0	-940.99	357.45	39	< .001 ***

The CLMM parameter estimates, standard errors, *z*-values, *p*-values, and odds ratios with 95% confidence intervals for each comparison are presented in Table B6.7. There were no significant interactions except for between Song and Story conditions at posttest, with an odds ratio of 0.02 (95% CI [0.00, 0.77], $z = -2.097$, $p = 0.036$), indicating a small effect over time for the Story group compared to the Song group. Whilst the actual results from the pilot were fairly meaningless for inferential purposes with only four participants per group and a redacted version of the intervention, it was possible to see that the data analysis procedures were going to produce appropriate output for running the CLMM and permit RQ2a to be addressed.

Table B6.7. CLMM Parameter Estimates and Odds Ratios (Pilot)

	Estimate	Std. Error	z value	Pr(>|z|)	2.50%	97.50%	Odds ratio
0|1	-2.5795555	0.49417016	-5.2199741	1.79E-07	-3.5481112	-1.6109997	0.0758077
1|2	-0.4003063	0.48188145	-0.8307153	0.4061345	-1.3447766	0.544164	0.67011477
2|3	1.63427281	0.48515973	3.36852526	0.00075572	0.68337721	2.5851684	5.12572924
3|4	2.96229209	0.49529446	5.98087059	2.22E-09	1.99153278	3.9330514	19.3422551
4|5	4.1500397	0.51243985	8.098589	5.56E-16	3.14567606	5.15440334	63.4365187
Song vs Chant	1.52591501	1.26247296	1.20867144	0.2267891	-0.9484865	4.00031655	4.59935009
Song vs Story	0.77192087	1.26344445	0.61096542	0.54122247	-1.7043848	3.2482265	2.16391888
Song vs Control	1.54282565	1.26935745	1.21543829	0.22419897	-0.9450692	4.03072054	4.6777894
T1 vs T2	0.57823974	0.66107077	0.87470171	0.38173623	-0.7174352	1.87391464	1.7828973
Syllables	-1.1086652	0.17521265	-6.327541	2.49E-10	-1.4520757	-0.7652547	0.32999914
Song vs Chant : T1 vs T2	-3.4122443	1.85239964	-1.842067	0.06546535	-7.0428809	0.21839228	0.03296713
Song vs Story : T1 vs T2	-3.9216944	1.87003461	-2.097124	0.0359826	-7.5868949	-0.2564939	0.0198075
Song vs Control : T1 vs T2	-1.4446257	1.84987448	-0.7809317	0.43484266	-5.070313	2.18106171	0.23583434

The data also permit the three remaining research questions to be addressed, by collecting information on whether the novel items of vocabulary were imitated (and to what extent produced correctly); and whether the grammatical errors were corrected or repeated verbatim.

B6.9 Summary of changes made after pilot

Having piloted the screening variables with 16 children, it was clear that the baseline tests (which also included the pretest) were taking too long with each child to be feasible once the number of participants increased to near 100. Therefore, the paper-based WASI test of verbal reasoning was replaced with the online PVST (Anthony & Nation, 2021) as described in 4.3.3.4, which meant that a whole group or class of children could be tested at once on this measure. Since English vocabulary knowledge was not the main focus of the experiment, the insight provided by the WASI vocabulary test into children's verbal reasoning seem unnecessarily detailed. All that was required is an indication of children's receptive vocabulary as a proxy measure of verbal ability. It was decided that the Matrix Reasoning test could be administered on paper as a quiet self-test exercise that children completed by circling in pencil their choice, as a class or group exercise rather than one-on-one. This would also help reduce the time taken to administer the screening variables to fit with the timeframe that schools were able to make available. The final refinement was to ask

all children to wear the headphones whilst completing the EIT, rather than giving them a choice to listen through the laptop speakers, as then the sound quality was more stable for all participants.

Appendix B7: Certificate of participation

Appendix B8: Data analysis

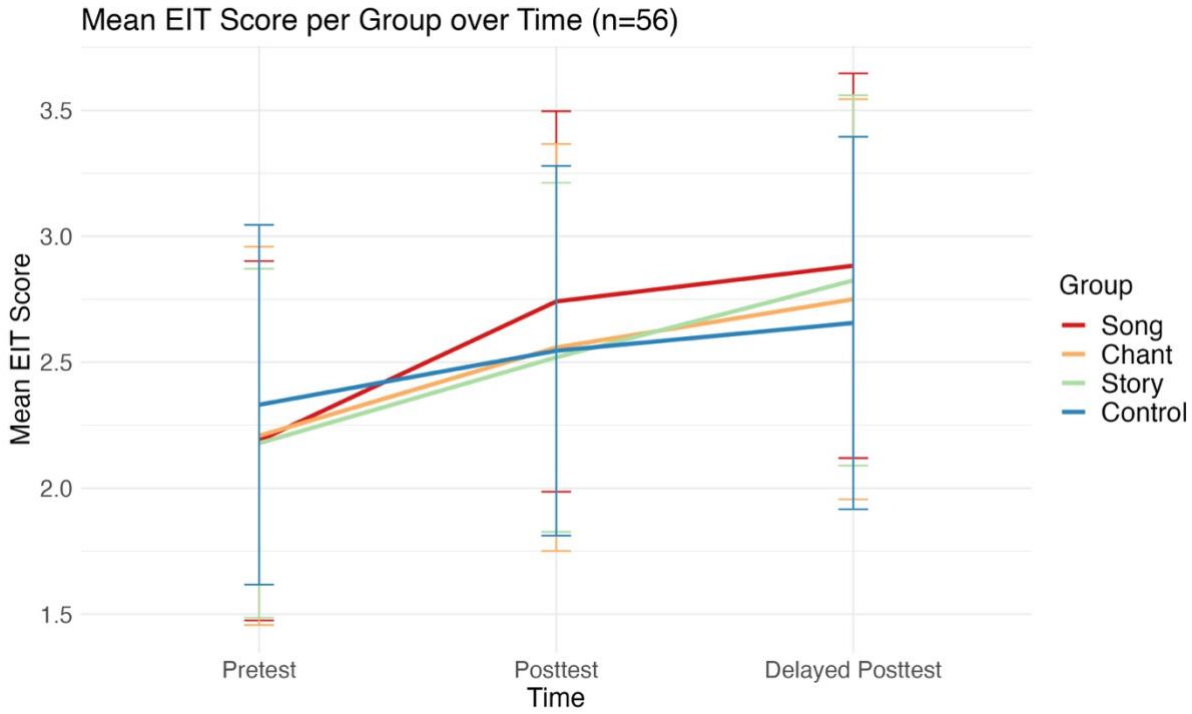
B8.1: Sensitivity analysis

I conducted a sensitivity analysis by running the CLMM and including only the 56 participants (n = 14 per group) who had attended every intervention lesson. Mean scores and standard deviations for each group are shown in Table B8.1. Figure B8.1 shows the group mean scores and 95% confidence intervals over time.

Table B8.1. Descriptive statistics for the sensitivity analysis

Group	Time	Mean	SD	SE	Lower 95% CI	Upper 95% CI
Song	Pretest	2.19	1.36	0.28	1.48	2.90
	Posttest	2.74	1.44	0.30	1.99	3.50
	Delayed posttest	2.88	1.46	0.30	2.12	3.65
Chant	Pretest	2.21	1.43	0.30	1.46	2.96
	Posttest	2.56	1.54	0.32	1.75	3.36
	Delayed posttest	2.75	1.52	0.32	1.96	3.54
Story	Pretest	2.18	1.32	0.28	1.49	2.87
	Posttest	2.52	1.32	0.28	1.83	3.21
	Delayed posttest	2.82	1.40	0.29	2.09	3.56
Control	Pretest	2.33	1.36	0.28	1.62	3.05
	Posttest	2.55	1.40	0.29	1.81	3.28
	Delayed posttest	2.66	1.41	0.29	1.92	3.40

Figure B8.1. Line plot of the mean EIT score per group (with 95% CI)



Results of the CLMM

There was a main effect of time, with all groups improving their scores at posttest (b = 0.74, SE = 0.15, z = 5.05, p < .001) and delayed posttest (b = 1.16, SE = 0.14, z = 8.11, p < .001). There was a main effect of syllables, with scores decreasing across all groups as the EIT stimuli increased in length (b = -1.41, SE = 0.16, z = -8.64, p < .001).

Experimental conditions compared to control

All experimental conditions scored more highly than the control group at posttest and delayed posttest. The interactions between the song and control condition over time from pretest to posttest (b = 0.71, SE = 0.34, z = 2.07, OR = 2.03, p = .038) and pretest to delayed posttest (b = 0.85, SE = 0.32, z = 2.66, OR = 2.34, p = .008) were statistically significant. This indicates that the likelihood of moving to a higher EIT score category from pretest to posttest was significantly greater for participants in the song condition compared to those in the control condition. The odds of score increase for the song condition were around twice those of the control over this period, suggesting that participants in the song condition showed greater improvement in performance from pretest to posttests. There was a significant interaction between story and control at delayed posttest (b = 0.71, SE = 0.31, z = 2.25, OR = 2.04, p = .024), but not at posttest (b = 0.38, SE = 0.34, z = 1.14, OR = 1.47, p = .25). The chant and control groups had no significant interactions at posttest (b = 0.23, SE = 0.34, z = 0.67, OR = 1.25, p = .51) or delayed posttest (b = 0.496, SE = 0.31, z = 1.62, OR = 1.64, p = .11).

Relative effects between experimental conditions

Similar to the main analysis, contrasts between experimental conditions indicated a decrease in likelihood of the chant condition (b = -0.48, SE = 0.33, z = -1.44, OR = 0.62, p = .15), and story condition (b = -0.32, SE = 0.32, z = -0.1, OR = .73, p = 32) participants scoring more highly than the song condition participants at posttest, and again at delayed posttest for chant (b = -0.36, SE = 0.32, z = -1.10, OR = .70, p = 0.27) and story (b = -0.14, SE = 0.31, z = -0.45, OR = .87, p = 0.66). However, none of these differences were statistically significant.

Since these findings reflected the same pattern as the analysis with 94 participants, I ran the CLMM with all participants as an 'intention to treat' analysis and report the findings from that analysis as my main findings.

B8.2: Model testing

Table B8.2 provides the output for the alternative CLMM that accounts for the nested data structure by adding in a Group/ID term to the first random effect for the contrasts with Control as reference. The model is coded as:

```
fmm2_nest <- clmm(EIT_score ~ (Control_VERSUS_Song + Control_VERSUS_Chant
+ Control_VERSUS_Story) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | Group/ID) +
(1 + (Control_VERSUS_Song + Control_VERSUS_Chant + Control_VERSUS_Story) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

Table B8.2a. CLMM output with Group/ID random effect and Control as reference group

	Estimate	Std. Error	z value	Pr(>|z|)	Lower CI	Upper CI	Odds ratio
0|1	-6.9063921	0.27643445	-24.983833	9.163E-138	-7.4482037	-6.3645806	0.00100136
1|2	-1.1098916	0.20991096	-5.2874401	1.2404E-07	-1.5213171	-0.6984661	0.32959468
2|3	0.47358031	0.20960526	2.25939133	0.02385905	0.062754	0.88440662	1.60573294
3|4	1.89986208	0.21100895	9.00370378	2.1823E-19	1.48628454	2.31343962	6.6849724
4|5	3.22516191	0.21410565	15.0634134	2.8185E-51	2.80551484	3.64480899	25.1576472
Control vs Song	-0.0939157	0.41998802	-0.2236152	0.82305676	-0.9170922	0.72926082	0.91035951
Control vs Chant	0.04450549	0.41273289	0.10783121	0.91412959	-0.764451	0.85346195	1.04551071
Control vs Story	0.08269812	0.4163163	0.19864252	0.84254239	-0.7332818	0.89867807	1.08621385
T1 vs T2	0.64647127	0.10822677	5.97330295	2.325E-09	0.43434681	0.85859574	1.90879332
T1 vs T3	0.99930064	0.10584401	9.44125818	3.6833E-21	0.79184637	1.2067549	2.71638143
Syllables	-1.4127081	0.13759062	-10.267473	9.8757E-25	-1.6823857	-1.1430304	0.24348302
Control vs Song: T1 vs T2	0.51883266	0.23052713	2.25063599	0.0244086	0.06699948	0.97066584	1.6800653
Control vs Song: T1 vs T3	0.71807224	0.24352008	2.94871881	0.00319094	0.24077288	1.1953716	2.05047658
Control vs Chant: T1 vs T2	0.36415592	0.22845372	1.59400303	0.11093536	-0.0836134	0.81192522	1.43929862
Control vs Chant: T1 vs T3	0.56829289	0.22689831	2.50461489	0.01225848	0.1235722	1.01301359	1.76525101
Control vs Story: T1 vs T22	0.42100942	0.24991471	1.68461242	0.09206337	-0.0688234	0.91084224	1.52349863
Control vs Story: T1 vs T3	0.59049217	0.24331806	2.42683253	0.01523128	0.11358878	1.06739556	1.80487651

And then for the contrasts with Song as reference, the alternative model was coded:

```
fmm3_nest <- clmm(EIT_score ~ (Song_VERSUS_Chant + Song_VERSUS_Story +
Song_VERSUS_Control) * (Time1_VERSUS_Time2 + Time1_VERSUS_Time3)
             + syllablesen2 +
              (1 + Time1_VERSUS_Time2 + Time1_VERSUS_Time3 | Group/ID) +
(1 + (Song_VERSUS_Chant + Song_VERSUS_Story + Song_VERSUS_Control) *
(Time1_VERSUS_Time2 + Time1_VERSUS_Time3) | Sentence.ID),
              data = df,
      link = "logit")
```

The output for the CLMM with Song as reference group is shown in Table B8.2b.

Table B8.2b. CLMM output with Group/ID random effect and Song as reference group

	Estimate	Std. Error	z value	Pr(>|z|)	Lower CI	Upper CI	Odds ratio
0|1	-6.9068951	0.27644422	-24.98477	8.951E-138	-7.4487258	-6.3650645	0.00100086
1|2	-1.1102717	0.20991256	-5.2892106	1.2285E-07	-1.5217003	-0.6988431	0.32946943
2|3	0.4731937	0.20960591	2.25753988	0.02397436	0.06236612	0.88402129	1.60511227
3|4	1.89949978	0.21100862	9.00200083	2.2164E-19	1.48592288	2.31307667	6.68255083
4|5	3.2247952	0.21410457	15.0617763	2.8892E-51	2.80515024	3.64444017	25.1484233
Song vs Chant	0.13799482	0.41620387	0.33155582	0.74022469	-0.6777648	0.95375441	1.1479696
Song vs Story	0.17627383	0.4207778	0.41892379	0.67527183	-0.6484507	1.00099831	1.19276463
Song vs Control	0.09366598	0.41997945	0.22302515	0.82351593	-0.7294937	0.91682569	1.09819287
T1 vs T2	0.64663267	0.10823325	5.97443668	2.3089E-09	0.43449551	0.85876983	1.90910142
T1 vs T3	0.99943866	0.10585319	9.44174365	3.6663E-21	0.79196641	1.20691091	2.71675638
Syllables	-1.4126583	0.13759151	-10.267046	9.9196E-25	-1.6823376	-1.1429789	0.24349514
Song vs Chant: T1 vs T2	-0.1547031	0.24093046	-0.6421069	0.52080377	-0.6269268	0.31752059	0.85666947
Song vs Chant: T1 vs T3	-0.1497584	0.24074655	-0.6220583	0.53390355	-0.6216216	0.32210486	0.86091596
Song vs Story: T1 vs T2	-0.0979229	0.24824731	-0.3944569	0.69324377	-0.5844876	0.38864187	0.90671885
Song vs Story: T1 vs T3	-0.1276312	0.23369193	-0.5461517	0.58496169	-0.5856674	0.33040495	0.8801779
Song vs Control: T1 vs T2	-0.5191354	0.23053405	-2.2518817	0.02432975	-0.9709821	-0.0672887	0.59503479
Song vs Control: T1 vs T3	-0.7183831	0.24354052	-2.9497479	0.00318033	-1.1957225	-0.2410437	0.48753991

Appendix B9: Verbal assent script

ASSENT SCRIPT TO BE READ BEFORE ASSESSMENT SESSION 1

I am a teacher and I want to know more about how children learn French. If you decide to help, I will ask you to do some little puzzles and point to words when I say them aloud, and tap along to a rhythm game that plays on the laptop. It is okay if you do not know the answers to the puzzles. I am just interested in hearing what you think and say, and having fun with the rhythm game.

I will also ask you to listen to a sentence in French, and then try to repeat it as best you can. I would like to record your voice for this part, if that's okay with you? We will use this microphone and computer. The recording will help me later when I'm trying to remember exactly what you said.

Then I'm going to teach you and your classmates French every day for a few weeks, doing lots of fun activities together. After that, we will do the French sentences game again where I say something, and you copy me. And then a few weeks later I'll come back to visit, and we will do the French sentences again if you are happy to join me.

If at any time you don't want to do any more activities with me, you can just say you want to stop, and we will stop straightaway. No one will be upset with you if you want to stop. It will be okay if you want to stop at any time or have a little break and try again later. You can just let me know, okay?

ASSENT SCRIPT TO BE READ BEFORE ASSESSMENT SESSION 2

Hello. Thanks for joining me again. Remember that you can just say if you want to stop, at any time. No one will be cross with you if you want to stop. You can just ask to stop completely or have a little break at any point, okay?

Appendix B10: CUREC approval email

Dear Catherine Hamilton,

Research ethics approval

Research title: Investigating the efficacy of whole-class singing activities for linguistic and educational outcomes of young language learners in UK primary schools.

Research ethics reference: C1A-23-015

The above application has been considered on behalf of the Education Departmental Research Ethics Committee (DREC) in accordance with the University's procedures for ethical approval of all research involving human participants.

I am pleased to confirm that, on the basis of the information provided to the DREC, ethics approval has now been granted for this study. Please find aproaval letter attached.

Please note the following:

Personal data: It is the responsibility of the PI to ensure that all personal data collected during the project is managed in accordance with the University's guidance and legal requirements.
In-person activities: Any data collection involving in-person interactions with participants must have an up-to-date fieldwork risk assessment in place; further guidance is available from the Safety Office's website.
Amendments: Please notify the committee if you intend to make any amendments to the information in your ethics application as submitted at date of this approval, as all changes must receive ethical approval prior to implementation. The amendment form is available on the SSH IDREC webpage.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to staff.curec@education.ox.ac.uk / student.curec@education.ox.ac.uk or ethics@socsci.ox.ac.uk.

Yours sincerely

Katharina Ereky-Stevens
DREC member

Appendix B11: Intervention materials

B11a: Song and chant condition materials

Song and chant condition input materials are identical, except the presence of the melody in the song condition and absence of the melody in the chant condition.

Song/chant title	Song/rhyme lyrics
Dans la forêt lointaine	Dans la forêt lointaine, on entend le coucou Du haut de son grand chêne, il répond au hibou: Coucou, hibou, coucou hibou, coucou hibou, coucou Coucou, hibou, coucou hibou, coucou hibou, coucou
Dansons la capucine	Dansons la capucine, y a pas de pain chez nous Y en a chez la voisine mais ce n'est pas pour nous. Youuu.
Dodo, l'enfant do	Dodo, l'enfant do, l'enfant dormira bien vite Dodo, l'enfant do, l'enfant dormira bientôt
Il était un petit navire	Il était un petit navire, il était un petit navire Qui n'avait ja-ja-jamais navigé, qui n'avait ja-ja-jamais navigé Oé oé
Jean Petit qui danse	Jean Petit qui danse, Jean Petit qui danse, De son pied il danse, de son pied il danse, De son pied, pied, pied. Ainsi danse Jean Petit.
La petite bête	C'est la petite bête qui monte qui monte C'est la petite bête qui monte qui monte C'est la petite bête qui monte qui monte Et jusqu'où? Jusqu'au cou! Guilli guilli guilli.
Meunier, tu dors	Meunier, tu dors. Ton moulin va trop vite. Meunier, tu dors, ton moulin va trop fort. Ton moulin, ton moulin va trop vite Ton moulin, ton moulin va trop fort Ton moulin, ton moulin va trop vite Ton moulin, ton moulin va trop fort Meunier, tu dors.
Mon petit lapin a bien du chagrin	Mon petit lapin a bien du chagrin Il ne saute plus, il ne danse plus Saute saute saute mon petit lapin Saute saute saute tu auras du thym
Petit escargot	Petit escargot porte sur son dos, sa maisonette Aussitôt qu'il pleut il est tout heureux Il sort sa tête
Petit poisson qui tourne en rond	Petit poisson qui tourne en rond Petit poisson qui n'a pas de nom Petit poisson rouge, petit poisson qui bouge Petit poisson qui tourne en rond, dis-moi ton nom

Promenons-nous dans les bois	Promenons-nous dans les bois Pendant que le loup n'y est pas Si le loup y était il nous mangerait Mais comme il n'y est pas il nous mangera pas Loup y'es-tu? Que fais-tu? J'arrive!
Savez-vous planter les choux?	Savez-vous planter les choux à la mode à la mode? Savez-vous planter les choux à la mode de chez nous? On les plante avec les pieds à la mode à la mode On les plante avec les pieds à la mode de chez nous.
Sur le pont d'Avignon	Sur le pont d'Avignon, on y danse on y danse Sur le pont d'Avignon on y danse tous en rond. Les éléphants font comme ça Et puis encore comme ça
Tourne, tourne, petit moulin	Tourne tourne, petit moulin Tapent tapent petites mains Nage nage petit poisson Vole vole p'tit papillon
Un kilomètre à pied	Un kilomètre à pied ça use, ça use Un kilomètre à pied ça use les souliers Deux kilomètres à pied ça use, ça use Deux kilomètres à pied ça use les souliers
Alouette, gentille alouette	Alouette, gentille alouette, alouette, gentille alouette Je te plumerai la tête, je te plumerai la tête Et la tête et la tête alouette alouette Ah ah ah ah Alouette, gentille alouette, alouette, je te plumerai
Ainsi font, font, font	Ainsi font font font les petites marionettes Ainsi font font font trois petits tours et puis s'en vont
Un éléphant se balançait	Un éléphant se balançait sur une toile toile d'araignée Il trouva ça si amusant qu'il appela un autre éléphant
Scions scions scions du bois	Scions scions scions du bois Pour la mère pour la mère Scions scions scions du bois Pour la mère Nicolas
Pirouette Cacahuète	Il était un petit homme, Pirouette cacahuète Il était un petit homme Qui avait une drôle de maison Qui avait une drôle de maison
Frère Jacques	Frère Jacques Frère Jacques Dormez-vous? Dormez-vous? Sonnez les matines. Sonnez les matines. Din dan don. Din dan don.

Rame rame rame donc

Rame rame rame donc vogue le canot
Joliment joliment joliment
Attaquons les flots

Une poule sur un mur

Une poule sur un mur
Qui picote du pain dur
Picoti, picota, lève la queue
Et puis s'en va

L'araignée Gipsy

L'araignée Gipsy monte à la gouttière
Tiens voilà la pluie
Gipsy tombe par terre
Mais le soleil a chassé la pluie
Et l'araignée Gipsy monte à la gouttière

B11b: Story condition materials

The story condition is a 12-installment story that is matched for lexical and grammatical complexity with the song and chant conditions.

Story installment	Input
1	Dans la forêt lointaine, il était un lapin, un hibou et un coucou. "Coucou. Coucou," dit le coucou du haut de son grand chêne. "Promenons-nous dans les bois?" "Jusqu'où?" "Jusqu'au Jean Petit qui danse." "Si le loup y était, il nous mangerait." "Le loup n'y est pas. Il nous mangera pas. On y va."
2	Au bois il y a un escargot qui porte sa maisonette sur son dos. La pluie! La pluie! Aussitôt qu'il pleut, il est tout heureux. Il sort sa tête, ainsi. "Coucou." Petit Lapin ne répond pas. Il y a une petite bête, qui monte, qui monte, qui monte. La petite bête monte bien vite au cou. "Coucou! Guilli, guilli, guilli!" Petit Lapin saute. Il ne danse plus. Ils font un kilomètre à pied. Deux kilomètres à pied. Ça use. Ça use les souliers.
3	Chez le Meunier, ce moulin. Mais le Meunier dors. Son moulin va trop vite. Son moulin va trop fort. "Dormez-vous? Ton moulin! Ton moulin! Ton moulin tourne bien trop vite! Ton moulin! TON MOULIN! Trop fort, ton moulin! Ton moulin tourne vite! Ton moulin va bien trop fort!" Le meunier dors. Ils s'en vont chez la voisine.
4	C'est Frère Jacques. "Coucou." Mais Frère Jacques dort. "Oé, oé! Dormez-vous? Sonnez les matines! Din dan don! Sonnez! Y es-tu? Tu dors?" Tiens! P'tit papillon qui vole sur la tête, Hibou! "Saute Petit Lapin – ainsi: saute, saute, saute. Tu auras du thym!" Il ne saute plus.
5	Ils arrivent à la drôle de maison de Pirouette Cacahuète. "Savez-vous planter les choux à la mode de chez nous, Hibou?" "On les plante avec la main à la mode?" "On les plante avec les pieds à la mode – ainsi, Hibou." "Ton pied? Pour planter choux?! Ton pied? Ça use! Jamais pour nous! Pieds!"
6	Il était un petit navire. "Rame donc, petit navire. Rame! Attaquons les flots! Vogue le canot! Rame!" Hibou n'avait jamais navigé. Il y a un petit poisson rouge qui tourne joliment en rond. "Dis-moi ton nom, petit poisson qui nage." Petit Poisson n'a pas de nom. Petit Poisson nage joliment. Poisson tourne joliment en rond, puis Poisson s'en va.

7 À un kilomètre, la gentille Mère Nicolas.
"Que fais-tu, la mère?"
"Scions. Scions, scions du bois, Hibou. Pendant que l'enfant fait dodo, scions, scions, scions. L'enfant était chassé. Do, mon petit enfant do."
"Qui ça?"
"L'enfant."
"Mon Petit Lapin dormira. Savez-vous danser, la gentille mère?"
"J'arrive bientôt."

8 Sur le mur, il y a une poule qui picote du pain dur et une gentille alouette qui vole.
"Picoti, picota. Alouette! Va t'en, Alouette! Gentille Alouette, je te plumerai la tête! Je te plumerai la tête, Alouette! Je te plumerai, Alouette!"
Alouette monte à la gouttière – la gouttière.
"Le pain n'est pas pour nous, Hibou."

9 Après, c'est le Pont d'Avignon.
Voilà la maison de Jean Petit, qui danse de son pied, comme ça.
On entend les petites marionettes.
"Dansons comme Jean Petit!"
Ils font comme ça: la capucine, pour les éléphants sur le pont.
L'enfant danse. Et puis encore.
Mais Petit Lapin a du chagrin. Qu'il danse!

10 Deux kilomètres à pied, ça use les souliers.
Petit Lapin tombe par terre.
Il y a une toile d'araignée Gipsy.
Un éléphant, qui monte trop, se balançe sur la toile.
Il appella un autre éléphant, qui monte.
Il font trois petits tours. Petit Lapin trouva ça si amusant, lève la tête…la queue bouge, et il saute! Il danse!

11 Le soleil monte, et danse.
Les matines font din dan don. C'est drôle.
Le loup n'y est pas.
Qui danse?
On danse avec tous les gens qui tapent les petites mains et puis font dodo.

B11c: Intervention lesson plans

The content of each lesson is shown in the table. The story input refers to the instalment of the story (see B11b for the content). The control input followed the teaching materials provided by the schools adapted from the Primary Languages Network.

	Song/chant input	Story input	Control input
Week 1			
Lesson 1	• Mon petit lapin • Dans la forêt lointaine	1	Greetings
Lesson 2	• Il était un petit navire • Dodo, l'enfant do • Petit poisson qui tourne en rond	2	What's your name?
Lesson 3	• Jean Petit qui danse • La petite bête	3	How are you?
Lesson 4	• Meunier, tu dors • Petit escargot	4	Count to 10
Week 2			
Lesson 5	• Promenons-nous dans les bois • Savez-vous planter les choux?	5	Colours
Lesson 6	• Sur le pont d'Avignon • Tourne, tourne, petit moulin	6	Animals intro
Lesson 7	• Un kilomètre à pied • Alouette, gentille alouette	7	What animal is it?
Lesson 8	• Ainsi font, font, font • Frère Jacques • Dansons la capucine	8	My favourite animal
Week 3			
Lesson 9	• L'araignée Gipsy • Rame rame rame donc	9	Days of the week
Lesson 10	• Scions scions scions du bois • Une poule sur un mur	10	What day is it?
Lesson 11	• Pirouette Cacahuète • Un éléphant se balançait	11	Months of the year