Education
Endowment
Foundation

Foreign language learning and its impact on wider academic outcomes: A rapid evidence assessment

Victoria A. Murphy, Henriette Arndt, Jessica Briggs Baffoe-Djan, Hamish Chalmers, Ernesto Macaro, Heath Rose, Robert Vanderplank & Robert Woore

University of Oxford
Department of Education

July 2020

Executive Summary

Background to the report

The Education Endowment Foundation (EEF) commissioned this Rapid Evidence Assessment (REA) with a view to understanding what is known from the research literature concerning learning a foreign language (FL) and its impact on students' wider academic outcomes. The key questions addressed examine: i) the research identifying what approaches to teaching FLs are being used and what variables impact on the effectiveness of these approaches; ii) the research which has examined the influence of learning a FL (or knowing another language) on other aspects of attainment; and iii) the impact of using a non-native language as the Medium of Instruction (MoI) on language learning and academic attainment. To address these questions, this REA first identified relevant extant synthesis literature (systematic reviews and/or meta-analyses) which we then updated by locating and reviewing more recent research that adopted either a Randomised Controlled Design (RCT) or a Quasi-Experimental Design (QED) as these research designs enable the determination of underlying causal relationships between variables.

The review questions posed in this REA are fundamental to our understanding of Language in Education policies and attainment. The research discussed in this report will lead to a greater understanding of the key findings in the literature, together with the trustworthiness of these findings, and will in turn enable strategic decision making regarding future research programmes, funding for research, and policy making.

Methods used in the review

The research questions posed in this review are vast, encompassing decades' worth of research examining numerous different characteristics and variables of Foreign Language (FL) and Second Language (L2) programmes. As this project is a REA, and hence there was a limited amount of time, the research team adopted a two-phase approach in carrying out the review. Many of the members of the review team themselves have been actively researching in this area and were consequently aware that there were some pre-existing reviews addressing some of the review questions posed in this REA. Therefore, in Phase 1 of this project we identified potential 'seed' reviews, REAs, systematic reviews or meta-analyses which had already been carried out which would provide a basis upon which to carry out further work. In Phase 1 a rigorous systematic approach was taken (see Section 2) where search terms were carefully chosen and applied to a range of relevant databases to uncover appropriate outputs. The search terms clarified the nature of the publication (systematic review or meta-analyses), context of work (foreign language teaching and learning or MoI contexts) and outcomes (language outcomes or wider academic attainment). A set of pre-specified inclusion and exclusion criteria enabled us to identify the relevant outputs, and all abstracts and summaries were double-screened by members of the review team. The quality and relevance of the systematic reviews, meta-analyses and REAs we examined in Phase 1 was assessed using an adapted version of the Critical Appraisal Skills Programme (CASP) Systematic Review Checklist with inter-rater reliability used to establish internal consistency in the application of this approach. Carrying out this procedure led us to identify six seed reviews which had been carried out in the last ten years to serve as the underlying basis for our updates. These seed reviews are: Fitzpatrick Morris, Clark, Needs, Tanguay and Tovey (2019); Fox, Corretjer, Webb and Tian (2019); Fox, Corretjer and Webb (2019); Goris, Denessen and Verhoeven (2019); Graham, Choi, Davoodi, Razmeh and Dixon (2018); Harris and Ó'Duibhir (2011); and Lo and Lo (2014).

The focus of Phase 2 was to update the selected seed reviews from Phase 1 using the same methods applied in the original respective reviews. This process involved conducting database searches using the same search strategies as in the original seed reviews, but which covered the time period between publication of the original review to 2019. A set of inclusion/exclusion criteria (the same ones used in the original seed reviews) were applied to the outputs of these searches, in addition to a Risk of Bias assessment. Primary research published since the publication date of the original seed reviews was then located and screened. In order to assess the quality of research outputs in Phase 2 we applied Gorard's sieve (Gorard, 2014) which enabled us to carefully reflect on the trustworthiness of the research findings. The application of this sieve was applied double-blind by a minimum of two members of the research team. In all cases throughout this methodology we achieved very high inter-rater reliability estimates. The quality of evidence of the respective updates is incorporated into our narrative synthesis of our findings in Section 3.

Findings

In Section 3.1 we discuss the findings of review questions one and three. These relate to approaches used to teach FLs and their effectiveness, together with the programme characteristics that impact on successful FL learning and teaching. A key imitation to be borne in mind is that the context of primary and secondary MFL teaching in England inevitably limits the relevance of much of the evidence reported in this REA. Many of the studies, especially those involving EFL, are set in contexts where there is not only both easy and expected access to English language sources but also an understanding that English knowledge and skills will be a valuable asset to learners in their future lives and careers. The two seed reviews synthesised and then updated in Section 3.1 were Fitzpatrick et al (2018) and Harris and Ó'Duibhir (2011). We found 21 new studies that addressed RQs 1 and 3 from the Fitzpatrick et al (2018) seed review and 8 new studies from our updates to Harris and Ó'Duibhir (2011). These updated studies are wide ranging in themes, covering such areas as the value of same language subtitles (captions) in the classroom for enhancing listening comprehension and vocabulary acquisition, phonological training, different forms of input and input processing for grammar development, task types and group activities in planning for presentations, intensive and extensive reading for vocabulary development, and different approaches to developing writing skills.

The general findings from this research indicate that more important than the specific method used is the way in which it is delivered and by whom. In other words, the programme characteristics and practitioner skills are key in impacting on successful FL learning and teaching. In general, approaches that are largely meaning-oriented, providing rich, authentic, and stimulating FL input for students, which increases the involvement load (how engaged the learner is with the task/language) tends to be more successful. At the same time, however, there are numerous studies indicating that within this meaning-oriented approach there is scope for careful attention to specific linguistics features - often referred to as Focus on Form (FonF). These approaches need to be strategically employed and their effectiveness very much depends on characteristics of both the teacher (in terms of their skills as a practitioner and their proficiency in the FL) as well as the proficiency of the learner.

The findings from these review questions also indicate a role for technology in supporting FL learning in classroom settings, but again, this needs to be carefully considered. Technology use just for the sake of it is not advantageous. However, judicious use of technology, video, film and TV (and their captions) can have a place in facilitating the development of foreign language knowledge and skill. In section 3.1 we break down the findings in relation to vocabulary and grammar, reading, writing, and speaking/listening skills. There are some common themes emerging from evidence here that as indicated above revolve around the importance of teacher proficiency, using rich and varied methods to provide students with experience of language through a variety of media, a careful transition from primary to secondary levels of education, the careful use of both implicit and explicit instructional approaches, appropriate use of technology, and strategy instruction.

In section 3.2 we summarise the findings that speak to the review questions concerning wider academic attainment of learning foreign languages. Our research indicated that there was no one seed review that we could use that spoke exclusively to the question of wider *academic* achievement in learning a FL. However, the two Fox et al (2019) reviews encompass this issue within a broad remit, to examine bilingual advantages as well as whether knowing and using another language confers benefits (beyond knowing the language). We found an additional 17 new studies in our update to the two Fox et al (2019) seed reviews. The methodologies included in these reviews were different from those updates to the other seed reviews. That is, many of the relevant studies did not employ an RCT or QED intervention design but rather adopted a standard quasi-experimental approach where the independent variable was whether the participants were bilingual (or not), and their performance was then compared (bilingual vs monolingual) on some dependent variable(s). As such, our application of Gorard's sieve (2014) did not fit neatly as the sieve was not set out to evaluate this type of research (i.e., designs other than interventions). Many of the studies in this section therefore received somewhat lower ratings than we would expect of good quality intervention research. These low ratings are somewhat artificial because many of these studies were examples of robust research designs and careful attention to methodological rigour, features that were not captured by the inelegant application of Gorard's sieve to nonRCTs (or intervention QEDs).

The general findings from the Fox et al (2019.1 and 2019.2) reviews, together with our updates indicate that a considerable amount of research has been carried out to investigate whether being bilingual confers cognitive advantages. Despite this wealth of research, the findings are mixed and hence at this point, a definitive conclusion is elusive. There is some evidence to suggest that there are positive impacts of FL learning on other developing knowledge in students, the most convincing of which relates to enhanced metalinguistic awareness, which is an important factor underpinning developing literacy skills. This is an area that demands a rigorous and systematic research agenda to more carefully ascertain wider impacts of FL learning.

In section 3.3 we review the evidence that speaks to the fourth and fifth review questions. These concerned the impact of using a non-native language as medium of instruction (MoI) on both language development and academic achievement. In carrying out this work we synthesised the findings of four seed reviews (Fitzpatrick et al, 2018; Graham et al, 2018; Goris et al, 2019; and Lo & Lo, 2014) and carried out respective updates. We found two new studies from the Fitzpatrick et al (2018) seed review, 8 new studies in our update to Graham et al (2018), 4 new studies in our update to Lo and Lo (2014) and no new studies in our update of Goris et al (2019). As with other review questions our analysis indicated mixed findings for the effectiveness of MoI programmes on language and academic attainment. Students' skills and proficiency is a key variable interacting with outcomes. While there is evidence that supports the implementation of MoI programmes for developing both language and content, the effectiveness of this approach interacts with many variables, and will depend on the settings, implementation strategies, skills of the practitioner, and individual learner characteristics. We also note that many of the mixed findings could be attributable to methodological variability across this sphere of research and we would encourage more careful research designs focusing on issues such as how participants are recruited for these studies, how individual differences impact on outcomes, closer examination of the pedagogical approach taken in MoI programmes, and longitudinal studies to examine the longer-term impact of learning within these educational settings.

Recommendations for further research

While it is now somewhat of a cliché for researchers to call for the need for further research, it is certainly a justifiable plea in the context of the review questions posed in this REA. There is a considerable amount of convincing research addressing RQs 1 and 3 which focus on FL learning. Much of this research is of a sufficient quality to be informative. However, even in this area there is a need for more work that systematically examines the interactions between key variables and language outcomes – and importantly, across different linguistic features, and different communicative skills.

A key focus of this REA was to examine the wider impact of FL learning on academic outcomes. We have demonstrated in this review that while there are a handful of studies examining this issue, this question is woefully under-researched. We understand the interest in examining whether there are cognitive advantages to being bilingual, but from an educational point of view, it is critical to examine whether and to what extent learning a FL in educational settings impacts on learning other content-related areas. Based on the work thus far, from research in this REA but also research in the extant literature that did not meet the criteria for inclusion in this review, we predict that there are likely to be many positive influences of FL learning on other aspects of educational attainment.

Finally, despite the global proliferation of MoI educational programmes such as English Medium of Instruction (EMI) and Content and Language integrated Learning (CLIL), the research on the effectiveness of these programmes is both scant and mixed. For every study demonstrating advantages for learning through an EMI context there may be others which question the value of this approach. We believe that again this equivocal picture of the impact of MoI on students' learning stems from the methodological variability inherent in the research thus far and we hope to see more systematic investigations in this area.

In summary, this REA presents an important foundation examining the extant literature which addresses fundamental questions relating to Language in Education programmes. We know that within educational settings Language underpins achievement across all areas. Children who have weak language and communication skills when beginning school are at risk of under-achievement (Whiteside, Gooch & Norbury, 2017) and learning a FL or participating within a MoI programme can impact on this important aspect of children's cognitive and academic development. We need more rigorous and systematic research examining all of the review questions posed in this REA and we anticipate that the analysis of research we provide of the extant literature in this document will offer a powerful catalyst to developing a rigorous and informative research agenda.

Contents

1. Introduction

The teaching and learning of foreign languages (FL) has been expanding throughout different educational contexts and different parts of the world for a number of years (Macaro, 2003; Macaro, Graham & Woore, 2015; Murphy, 2014; Murphy & Evangelou, 2015; Wivers, 2018). The reasons for this development are many and varied. In many countries, the foreign language of choice is English, a choice motivated by a perceived need for non-native speakers of English to develop sufficient English proficiency so as to provide greater opportunities in both education and employment. Other reasons for learning languages might stem directly from the linguistic landscape of different countries. For example, Canadian children who are native speakers of English will typically learn French as a second language in school since French is one of the two official languages of Canada. In other contexts, such as the UK a range of different FLs may be taught based on a belief that learning FLs conveys some advantages on students and can lead to enhanced opportunities. Whatever the reasons espoused, different language learning programmes are being increasingly offered through schools and/or universities throughout the world.

Language learning within education can be supported through numerous different educational programmes. The most frequent setting is the foreign language learning context where students spend some portion of their time each week in school studying a foreign language as a taught subject. These contexts are typically input-limited where students might spend anywhere from 30 minutes to 3 hours per week in classrooms learning the FL (Murphy, 2014). Another context in which students learn language through education is the case of minority language learners in majority language contexts. Here students from (typically) ethnic minority backgrounds who have a home language that is not the same as the wider societal language, and consequently not the language of education, are taught through the medium of the majority language. Students within these contexts have varied linguistic and academic outcomes and the extent to which they are successful can depend on numerous variables including support for the home language, oral language skills (in the first and second language), and pedagogical approach to name but a few (Murphy, 2018; 2019). This context is growing internationally as a function of migration, globalisation of commerce, and the international refugee crisis. Despite estimates suggesting there are millions of children around the world in this educational setting (Murphy, 2014) there is a worrying lack of research and evidence that directly speaks to how we can support these children's linguistic and academic outcomes (Murphy & Unthiah, 2015; Oxley & De Cat, 2019).

Another context of learning language through education stems from the immersion model where children are taught academic content through the medium of a language they are also developing. One of the earliest examples of this approach was that developed by Lambert and Tucker (1972) in Montréal, Canada to help support English-speaking children's proficiency in French. This French immersion model proved to be very successful in supporting both dual language and biliteracy skills at no cost to academic achievement (see Murphy, 2014 for a review). Dual immersion programmes have also been developed which aim to provide support for ethnic minority pupils. In the US, for example, where there is a high proportion of Spanish-speaking children in some states, children can attend programmes where part of their school day is spent in English (the majority language) and the other part is in Spanish[1]. They receive language arts instruction in both English and Spanish, as well as academic content taught through both English and Spanish. Importantly too, children in these contexts share the classroom with majority language speakers where the English-speaking children learn Spanish and the Spanish-speaking children learn English. These programmes, like the original French immersion model, have been successful in supporting children's language and academic outcomes, as well as supporting a positive self-image as bilingual speakers (see Murphy, 2014 for a review). However, most of the research evaluating the success of these programmes has been carried out in the context of North America (US and Canada). The success of these original models, however, has led to a global proliferation of medium of instruction (MoI) models. Arguably the most common of which is English Medium Instruction (EMI) models. EMI programmes are found in non English-speaking countries, at all levels of education (Early Childhood Education and Care (ECEC) through to Higher Education (HE)). Despite the global reach of EMI programmes, there is actually a relative lack of evidence which speaks to the success of these programmes (Macaro, 2018).

[1] These programmes are available in a variety of different languages in both the US and around the world.

Given the global reach of FL/L2 learning and teaching, and the fact that students in FL classrooms can be found at all stages of education, it is clearly important to have a solid understanding of factors which influence the learning of foreign and second languages, and consequences therein. In an effort to closely examine the current evidence base which speaks to the effectiveness of these different kinds of language in education programmes, the Education Endowment Foundation has commissioned this Rapid Evidence Assessment (REA) which addresses some of the most fundamental questions within the sphere of language education. In particular, this REA aims to examine the best evidence which speaks to the approaches taken to teach foreign languages, what variables contribute to the success of these approaches, what the wider impact of learning foreign languages might be on students, and whether content based (MoI) instructional models are supported by evidence in terms of students' language and academic outcomes.

The EEF's overall aim for this project is to understand the impact of foreign language learning and the most effective strategies to achieve language proficiency and positive impact on wider academic attainment. This evidence will be used to inform policy and practice, with the possibility of providing the basis for further primary research.

1.1 Objectives

The main objective of this evidence assessment is to understand what is known from the literature about learning a foreign language and its impact on students' wider academic outcomes. The specific review objectives are to summarise the evidence on:

- how to effectively teach a foreign language
- the effect of learning a foreign language on attainment in other academic subjects
- the effect on second language acquisition and on academic attainment of using a non-native language as the medium of instruction in academic subjects

and

- to provide practical recommendations on:
    - how to best teach a foreign language
    - how to best teach a foreign language to maximise benefits on wider academic outcomes
    - when and how to introduce a non-native language as the medium of instruction

1.2 Review questions

This rapid evidence assessment was guided by the following six review questions (RQs):

Primary questions

1. What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
2. What is the impact of learning a foreign language on students' wider academic outcomes?

Secondary questions

3. What practitioner skills or programme characteristics contribute to effective language learning among students?
4. What is the impact of using a non-native language as the medium of instruction in academic subjects on students' academic outcomes?
5. Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?
6. What is the impact of delaying or accelerating the introduction of a new 'local' language as a medium of instruction for new arrivals (e.g., refugees, immigrants) who are not yet proficient in their native language?

In addressing these review questions, we focused on research which adopted either a Randomised Controlled Trial (RCT) or quasi-experimental intervention design (QED). The RCT is considered by many to be a gold-standard in research as it enables the identification of causal relationships. Many advocate their increased use in understanding 'what works' in educational settings (e.g., Chalmers, 2018; Connelly, Briggart, Miller, O'Hare & Thurston, 2017). Consequently, research included in this REA were required to adhere to either RCTs or QEDs. There was one exception to this in section 3.2 where research is discussed from two seed reviews on research primarily following a quasi-experimental design but for studies that are not interventions. In these studies, the independent variable was typically 'bilingualism' – where bilinguals were compared against monolinguals on one or more dependent variables. While some of these studies were clearly neither RCTs and could not be considered educational interventions, they nonetheless adhere to a rigorous quasi-experimental design which speaks to the review question considering wider (academic) outcomes on knowing and using another language. As such, they were included in this REA.

These review questions are vast. A *rapid* assessment of evidence on such vast review questions is challenging to say the least. The strategy we adopted was to identify rigorous extant systematic reviews of the literature which addressed the key questions of this review.

We therefore undertook a carefully constructed search to identify those ‘seed’ reviews, which we then updated following their own methodologies. Our methodological protocol is provided in detail in Section 2. Sections 3.1, 3.2 and 3.3 present the detailed discussion of the outcome of our search for the seed reviews and our updates therein. Section 4 provides a synthesis of the research findings across the different updated reviews, and the REA concludes with a summary of recommendations for future work.

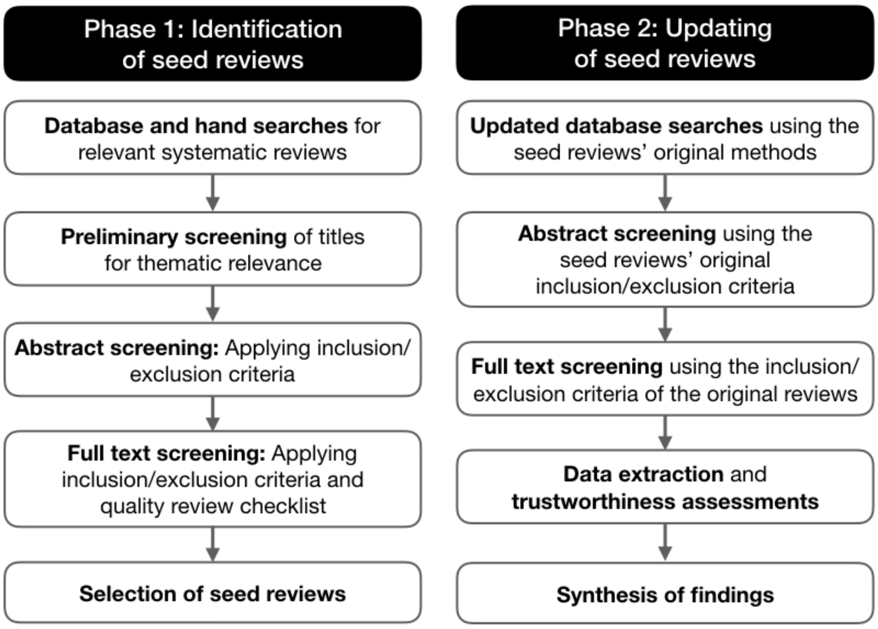
In summary, this REA provides a detailed analysis, within a single document, of the current state of the art of research which speaks to some of the most fundamental issues concerning language development in education. It will therefore serve as an important foundation for furthering our understanding of what approaches are proving to be successful in developing L2 skills, which programmes and programme characteristics are most likely to yield success, what are some of the wider consequences of knowing and using more than one language, and importantly what is yet needed to be done.

2. Methodology

2.1 Overview

Figure 2.1 provides an overview of the steps involved in gathering the research evidence that forms the basis of this REA. As stated in the introduction, given the broad nature of the review aims and objectives, and the short timescale, the project focused on appraising, synthesizing, and updating findings from previously conducted systematic reviews and meta-analyses.

Figure 2.1: Overview of REA methods


Phase 1: Identification of seed reviews
Database and hand searches for relevant systematic reviews
Preliminary screening of titles for thematic relevance
Abstract screening: Applying inclusion/ exclusion criteria
Full text screening: Applying inclusion/exclusion criteria and quality review checklist
Selection of seed reviews
Phase 2: Updating of seed reviews
Updated database searches using the seed reviews' original methods
Abstract screening using the seed reviews' original inclusion/exclusion criteria
Full text screening using the inclusion/ exclusion criteria of the original reviews
Data extraction and trustworthiness assessments
Synthesis of findings

Phase 1 consisted of a wide-reaching trawl for systematic reviews addressing the above RQs. These were assessed for relevance and quality, with the aim of selecting the most relevant and highest quality reviews as 'seed' reviews. Phase Two consisted of updating the seed reviews by replicating their methods, including search strategy and inclusion/exclusion criteria. In addition to these, inclusion was limited to papers published after the original search and to only RCTs and QEDs with a control group and pre- and post-tests. Any new studies meeting these inclusion criteria were incorporated into the findings of the original reviews. Finally, the findings of the seed reviews and papers identified in the updated search were analyzed thematically, across reviews, in a narrative synthesis.

2.2 Phase 1: Identifying and appraising existing research syntheses

The research team first conducted a broad search for systematic reviews which focused on one or more of the following topics covered by the review questions:

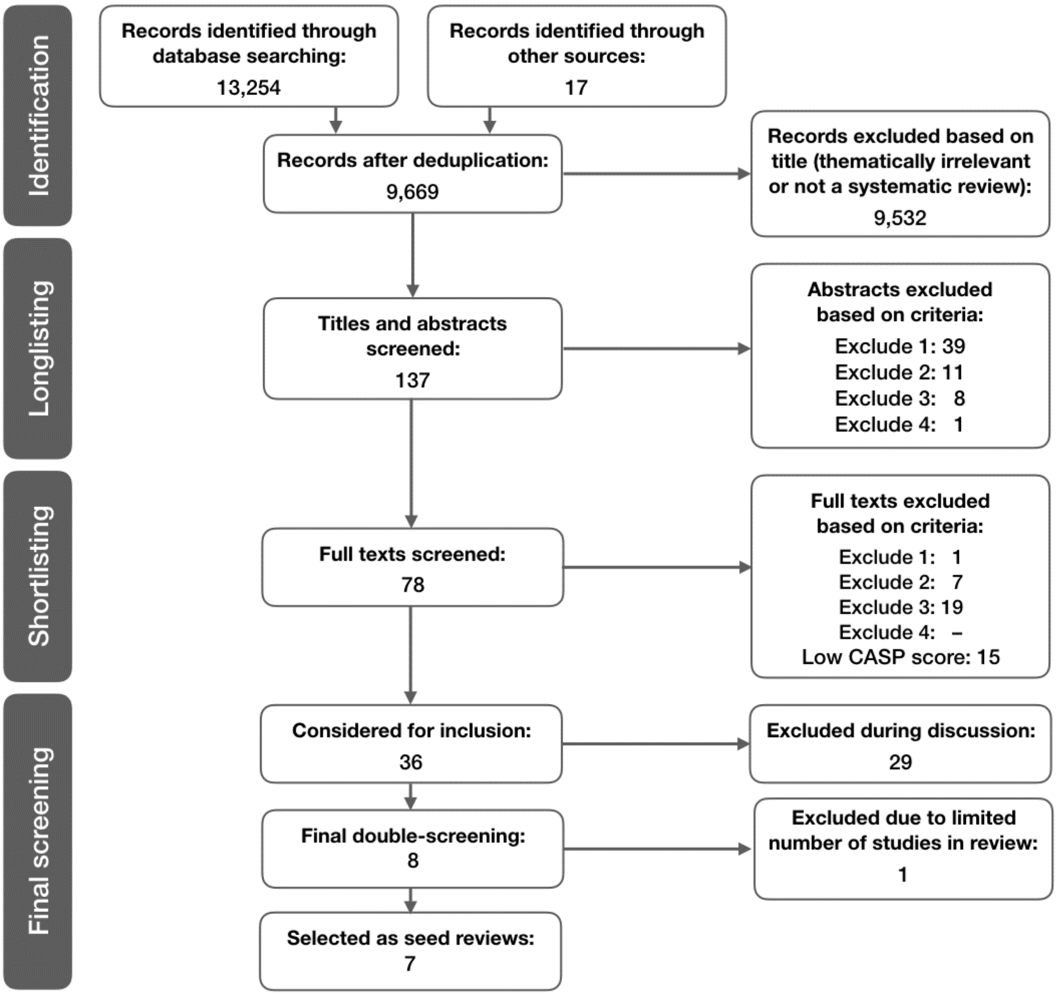
1. The effectiveness of different approaches to foreign language[2] teaching (RQs 1 and 3)
2. The wider personal and academic outcomes of foreign language learning (RQ 2)
3. The language and wider academic outcomes of using a foreign language as the medium of instruction (RQs 4 and 5)
4. The transition to majority language instruction for new arrivals, i.e. refugees or other immigrants (RQ 6)

Relevant syntheses were identified mainly through systematic electronic database searches (Section 2.2.1). In addition, the researchers asked colleagues and members of professional networks in the field of Language Education to provide information about any research syntheses addressing the focus areas of the REA (the text of this call and the list of networks to which it was sent is included in Appendix 1). Finally, hand searches were conducted of the online libraries of the EPPI-Centre (eppi.ioe.ac.uk) and Campbell Collaboration (www.campbellcollaboration.org), both of which are commissioners and publishers of systematic reviews in Education. The identified documents were subsequently screened for relevance to the aims of the REA (Section 2.2.2) and quality (Section 2.2.3). Finally, the researchers chose which seed reviews to update in the current REA from the pool of syntheses that were judged to be most relevant to the review objectives and of the highest quality (Section 2.2.4).

Figure 2.2 reports the overall number of records at each stage, from the initial identification throughout the subsequent screening process. The number of identified and screened records for each of the four aforementioned themes are reported in Appendix 2.

[2] Foreign language learning refers to the acquisition of a second or additional language which is not the majority language in the respective context.

Figure 2.2: Flow diagram of seed review screening and selection


Identification
Longlisting
Shortlisting
Final screening
Records identified through database searching: 13,254
Records identified through other sources: 17
Records after deduplication: 9,669
Records excluded based on title (thematically irrelevant or not a systematic review): 9,532
Titles and abstracts screened: 137
Abstracts excluded based on criteria:
Exclude 1: 39
Exclude 2: 11
Exclude 3: 8
Exclude 4: 1
Full texts screened: 78
Full texts excluded based on criteria:
Exclude 1: 1
Exclude 2: 7
Exclude 3: 19
Exclude 4: –
Low CASP score: 15
Considered for inclusion: 36
Excluded during discussion: 29
Final double-screening: 8
Excluded due to limited number of studies in review: 1
Selected as seed reviews: 7

2.2.1 Search terms and sources

The list of electronic databases for the systematic literature search, as well as the list of search terms used, were chosen in consultation with a research librarian at the Bodleian Education Library. Systematic searches were conducted of five databases deemed to be the most relevant to the aims of this REA: British Education Index, Education Collection (incl. ERIC), Linguistics Database (incl. LLBA), SCOPUS, and Web of Science. Four searches were conducted of each database, with slightly different search terms to address the different review questions driving the REA (see Appendix 3).

All searches included a set of terms specifying the type of publication sought (e.g. systematic review or meta-analysis). In addition, different terms were included in each search which clarified the context (e.g. general foreign language teaching and learning or L2 medium-of-instruction contexts) and the outcomes (e.g. language- or content-learning or wider academic outcomes) targeted by the relevant review questions. In all cases, the search terms clarifying the types of publications sought were limited to appearances in the document title only. All other terms were limited to appearances anywhere but in the main text (i.e. title, abstract, tags, and so on). Searches were also limited to documents published after 1999.

The searches were completed between September 23rd and 27th, 2019. The bibliographic information of the records returned by these searches (n=13,254) was exported and entered into an Excel spreadsheet, where it was merged with the information about the reviews identified through hand searches or via recommendations from professional networks (n=17) and all duplicate entries (n=3,585) were removed.

2.2.2 Longlisting: Assessing relevance

At this point, any publications were excluded which, according to the document title, were clearly not systematic reviews or meta-analyses or addressed topics that were not directly relevant to the current REA (e.g. concerning first language acquisition or learner populations with speech and language impairments). Where the relevance of a review could not be judged based on the title alone, the document was retained for the next stage of screening. Electronic copies were obtained of the 137 eligible documents which remained after the initial review and deduplication.

Next, we reviewed the abstracts, structured summaries and/or executive summaries of the longlisted publications, applying the inclusion and exclusion criteria listed in Table 2.1. Each of the documents marked for inclusion was also tagged with the review question or questions which the reviewers thought it helped to address. All abstracts and summaries were dual screened; that is, each was read by two members of the research team and any disagreements were resolved through discussion. In cases where a clear decision on any of the inclusion/exclusion criteria could not be reached based on the abstract or summary alone, the publication was retained for additional review. In total, 78 publications were longlisted for full-text review.

Table 2.1: Inclusion and exclusion criteria

Include	Exclude

Criterion 1: Publication type	The document is a systematic review or meta-analysis or refers to systematic methods used to locate and appraise the included literature.	The document is not a systematic review or does not refer to systematic methods used to locate and appraise the included literature.
Criterion 2: Context	The review is wholly or mainly focused on a school aged population of learners (4-18 years old) and has relevance to one or more of the following: (a) Teaching and learning foreign language in school settings, i.e. it was mainly engaged with pedagogy and classroom practice (b) Teaching and learning through the medium of the foreign language, not the majority language (c) Teaching and learning through the medium of the majority language in the case of new arrivals	The review is mainly based on studies of adult learner populations and/or does not have relevant either to (a) teaching and learning FL in school settings, (b) using the FL as the medium of instruction, or (c) using the majority language to instruct native speakers of other languages (e.g. English as an Additional Language). For example, the review may be wholly or mainly focused on incidental language learning or study abroad.
Criterion 3: Objectives	The review reports on the (relative) effectiveness of one or more of the following: (a) Different approaches to foreign language teaching (b) Using the foreign language as the medium of instruction, in terms of language and/or wider academic outcomes (c) Accelerating or delaying the instruction in the majority language for new arrivals	The review does not report on the (relative) effectiveness of (a) alternative approaches FL teaching, (b) FL medium of instruction, or (c) the time of transition to majority language instruction. For example, it is a review of theory, research methodology, learner and/or teacher perceptions or attitudes towards FL instruction, etc.
Criterion 4: Outcomes	The review reports on substantive educational outcomes such as test scores or exam pass rates, including if these are not directly related to FL outcomes, such as wider academic attainment.	The review reports only on non-educational outcomes such as student satisfaction or is purely descriptive.

2.2.3 *Shortlisting: Assessing quality*

Next, we conducted full text reviews of the 78 longlisted publications. In the first instance, the full text documents were judged against the same inclusion/exclusion criteria presented above. Seven syntheses (10%) were double screened by two team members each. The agreement rate exceeded 90%, which was considered sufficiently close to allow for the remainder of the studies to be appraised independently.

26 reviews were excluded at this stage because they did not meet one or more of these criteria. The remaining 51 syntheses underwent additional review using an adapted version of the Critical Appraisal Skills Programme (2018) Systematic Review Checklist (CASP; Appendix 4). The CASP checklist helped the researchers to assess the quality of the research syntheses in general methodological terms. It includes judging the appropriateness of the

methods used in the review, the likelihood that the review was exhaustive in relation to its stated aims, the quality of the literature informing the review (including the use of Risk of Bias or Weight of Evidence Assessments in the individual studies), and the quality and appropriateness of the methods used to synthesize the results. Six publications (c. 10%) were rated by two team members each to check for consistency in applying the CASP Checklist and any ambiguities were resolved through discussion among the whole team.

For each item on the CASP checklist, the reviewers assigned a score between 0 (denoting that the research synthesis did not meet the minimum expectations for this item), and 5 (reflecting that the synthesis met the highest expectations for this item). Based on these item scores, an average score was then calculated for each of the research syntheses. The reviewers also assigned an overall score (0–5) to each review, indicating whether it was of sufficient relevance and of high enough quality to be considered for inclusion in the REA (see adapted CASP checklist, item 12).

The scores assigned by the review team are listed in Appendix 5. Fifteen reviews, which had an overall score below 3, were excluded at this point for being of insufficient quality to serve as a basis for the REA. The remaining 36 syntheses were judged to be both sufficiently relevant and methodologically sound and were thus considered for inclusion as seed reviews.

2.2.4 Selecting seed reviews

The aforementioned quality assessment served as the basis for a discussion among the research team with the objective of selecting the highest quality syntheses which were also the most relevant to the four review themes specified at the top of Section 2.2. Since most reviews addressed only one of these themes, we aimed to select several reviews per theme.

The research team first considered those reviews which had been given the highest overall score (item 12) on the adapted CASP checklist, as well as the highest average score (mean of items 1–12). In addition, more recent reviews were given preference over older reviews and reviews which covered the review questions more thoroughly were given preference over those which covered them only partially. Furthermore, the reviews addressing the same theme were selected to be thematically complimentary: For example, the highest scoring review assessing the outcomes of FL medium of instruction was focused entirely on English-medium instruction in Hong Kong (Lo & Lo, 2014). Therefore, the researchers also selected syntheses which included studies conducted in other contexts (Goris, Denessen & Verhoeven, 2019; Graham, Choi, Davoodi, Razmeh & Dixon, 2018).

By this method, we selected two to three potential seed reviews per research topic. The aims and methods of these reviews are briefly summarised in Table 2.2, whereas their outcomes will be discussed in the findings section of this report. It must be noted at this point that no suitable systematic reviews were identified in this REA which addressed the fourth research theme (the transition to majority language instruction for new arrivals, RQ6)[3]. We therefore exclude RQ6 in our discussions henceforth.

As the relevance and quality of some of the studies selected as seed reviews had only been assessed by one researcher during the previous shortlisting phase, we conducted additional assessments at this stage using the same adapted CASP checklist. Each of the potential seed reviews was thus assessed by a total of three members of the research team, and all of them were agreed to be highly relevant and of acceptable quality to form the basis for this REA.

Table 2.2: Overview of systematic reviews and meta-analyses selected as seed reviews and which themes in the current REA they address

Themes	Aims and Methods
Fitzpatrick, Morris, Clark et al. (2018): Rapid evidence assessment: Effective second language teaching approaches and methods.	
(1) Approaches to FL teaching; (3) Outcomes of FL as medium of instruction	This review constitutes a Rapid Evidence Assessment, commissioned by the Welsh Government, to inform the provision of Welsh language teaching in primary and secondary schools. It was guided by the research question: ‘What teaching approaches and methods are effective in developing young learners’ second language competence according to high quality empirical evidence?’. The researchers systematically selected and synthesized 106 studies on the effectiveness of different approaches to language teaching and learning (including foreign language teaching, heritage language learning, and the use of the second language as the medium of instruction) and discussed of how their findings can be applied to the particular context of Welsh-language teaching.

[3] We believe the reason for failing to find suitable systematic reviews or meta-analyses to update in response to RQ6 stems from the fact that the most typical scenario for new arrivals is that they are immersed in the majority language in their respective educational context. Therefore, there is usually no facility to delay (or indeed accelerate) the introduction of the majority language as the student is typically admitted into the mainstream educational setting immediately upon arrival and taking up formal education. The proficiency of the child’s home language is equally not taken into consideration in these contexts. See Murphy (2014; 2018; 2019) for further discussion of the language development and academic achievement of ethnic minority pupils.

Harris & Ó Duibhir (2011): Effective language teaching: A synthesis of research.

(1) Approaches to FL teaching;
(3) Outcomes of FL as medium of instruction

This review was commissioned by the National Council for Curriculum Assessment in Ireland to inform the development of curriculum for language teaching in primary school. The researchers identified and synthesized 12 key studies focusing on practical approaches to teaching languages in the classroom with the aim of generalizing the findings to a set of key principles for successful language teaching that could be applied in the Irish context. Originally, the review considered only findings from studies with a process-product design (i.e. clear measures of effectiveness are linked to well defined and well measured instructional practices). Nevertheless, the report also includes a section on the outcomes of process-type and correlational studies which the authors identified as bearing strong relevance to the aims of the review. The findings of the reviewed studies are discussed with regard to their implications for classroom teaching, policy, and future research needs.

Fox, Corretjer, Webb & Tian (2019): Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2005–2011.

Fox, Corretjer & Webb (2019): Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2012–2019.

(2) Outcomes of FL learning

The purpose of this two-part systematic review was to provide a comprehensive survey and analysis of empirical research which shows the benefits of knowing more than one language (including foreign language learning, bilingualism, and multilingualism). The first review (Fox, Corretjer, Webb & Tian, 2019) synthesized findings from 65 studies published between the years of 2005 and 2011, whereas the second publication (Fox, Corretjer & Webb, 2019) covered 100 publications from the period between 2012 and 2019. With their review findings, the authors aimed to address the conception that bilingualism is negatively associated with intelligence and other competencies, which they note as being widespread in the United States among policy makers and members the general public. The systematic reviews were guided by two research questions: (1) 'What are the effects of foreign language/world language (FL/WL) learning and bilingualism on academic achievement, cognitive abilities, and learners' attitudes and beliefs?'; (2) 'What additional effects and factors may be associated with FL/WL learning or bilingualism drawn from the empirical research literature?'

Goris, Denessen & Verhoeven (2019): Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies.

(3) Outcomes of FL as medium of instruction

This systematic review focused on the effects of CLIL in the context of English-medium education in Europe. In particular, the authors searched for longitudinal studies of the effects of CLIL on students' language knowledge, including 'vocabulary, grammar, idioms and text comprehension' (p. 679). In discussing the results of their research, Goris et al. separately considered the outcomes of four studies of content and language integrated learning at the primary school level and eighteen studies conducted in secondary schools.

Graham, Choi, Davoodi, Razmeh & Dixon (2018): Language and content outcomes of CLIL and EMI: A systematic review.

(3) Outcomes of FL as medium of instruction

This study was focused on the content as well as language learning outcomes of English medium-instruction education and content and language integrated learning in countries where English is not the majority language. Unlike Goris et al. (2019), Graham and colleagues did not limit their search to research conducted in European schools. However, of the 25 studies identified in their systematic literature search, 23 were from Europe and only two had been conducted in Asia. In discussing their findings, the authors considered the

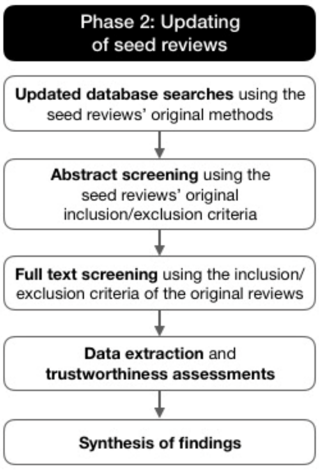
	development of productive and receptive language skills and also separately discussed the results from studies of English-medium Maths, Science, and tertiary education.
Lo & Lo (2014): A meta-analysis of the effectiveness of English-medium education in Hong Kong.	
(3) Outcomes of FL as medium of instruction	With their review of studies on English-medium instruction in Hong Kong, Lo and Lo (2014) aimed to provide data that could inform the country's future educational policy, as well as research comparing the effectiveness of EMI education in different context. The authors systematically selected and statistically synthesised the findings of 31 studies which compared the educational outcomes of students in secondary-level English- and Chinese-medium of instruction (EMI and CMI) programmes. Therein, they were guided by three research questions: (1) 'What is the difference in academic achievements between students studying in EMI and CMI education?'; (2) 'Are there differences in affective variables, including self-concept, motivation, learning strategies, and interest between students studying in EMI and CMI education?'; and (3) What are the variables which moderate the differences between students studying in EMI and CMI education (e.g. features of the research design or characteristics of the participants)?

2.3 Phase 2: Updating the seed reviews

In the second phase of the REA, the review team updated the selected seed reviews on the basis of the methodology used in the original. This involved conducting database searches using the same search strategies as the original reviews but covering the time period between the original searches and the REA. The identified records were reviewed using inclusion/exclusion criteria based on those used in the seed reviews, in addition to also undergoing a Risk of Bias assessment.

Figure 2.3 illustrates how the seed reviews were updated, from the initial database searches throughout the subsequent screening process. The number of identified and screened records for each of the six seed reviews are reported in Appendix 6.

Figure 2.3: Flow diagram of updated seed review articles screening and selection


Phase 2: Updating of seed reviews
Updated database searches using the seed reviews' original methods
Abstract screening using the seed reviews' original inclusion/exclusion criteria
Full text screening using the inclusion/ exclusion criteria of the original reviews
Data extraction and **trustworthiness assessments**
Synthesis of findings

2.3.1 Electronic database searches

Searches were conducted to locate primary research published since the date of the original search (where stated) or the date of publication of the seed review (where the search date was not reported). Only electronic databases were searched; that is, no handsearching was conducted due to the time constraints on this REA. The same search strategies as in the original seed reviews were used wherever possible (bibliographic databases, search terms, and other limitations listed in Appendix 7). Some searches could not be replicated, however, as the team did not have access to some of the databases consulted in the original reviews.

The updated searches were conducted between November 25th and December 3rd, 2019. The returned records were uploaded to Rayyan, a web-based application for collaborative abstract screening in the preparation of systematic reviews (Ouzzani et al. 2016). A separate Rayyan database was created for each seed review.

2.3.2 Abstract screening

In Rayyan, we reviewed the identified publications' titles and abstracts, applying the same inclusion/exclusion criteria as used in the seed reviews (summarized in Appendix 8), with the following addition: Only randomized control trials (RCTs) and studies with a quasi-experimental design (QEDs[4]) were considered for inclusion.

[4] We operationalized QEDs as formal comparisons in which alternative teaching approaches or conditions are evaluated against each other (i.e. a treatment and comparator/control). By way of illustration, we included non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. We excluded single group pre-post designs, case studies, ethnographies, and cross-sectional designs. Methodological shortcomings are reflected in the weight of evidence assessments for new studies in the update.

This additional criterion was applied to all seed reviews except the one conducted by Fox et al. (2019), which surveyed studies focusing on foreign language learning and bilingualism. Bilingualism is a variable over which experimenters can have little or no control, which makes RCTs impossible. There can be no experimental manipulation of the independent variable when the variable is etiologic to the participants. Therefore, for this seed review, the additional inclusion criterion was expanded to studies in which bilingual participants were compared to monolingual participants. These studies adhered to a quasi-experimental design in that there was no random allocation to groups since participants came to the study already bilingual (or not). In these studies, 'Bilingualism' is the grouping variable where bilingual performance on some (set of) dependent variable(s) is compared against monolinguals. Thus, studies which involved only bilingual participants and used, for example, regression analyses to assert bilingual advantages, were not included.

The first 10% of the records identified in the updated database searches from each seed review were screened by two team members and their decisions compared. In all cases, agreement was 90% or higher, which demonstrated consistency in applying the updated inclusion/exclusion criteria. Any records which, on the basis of the information contained in the titles and abstracts, did not meet one or more of the inclusion/exclusion criteria were tagged as 'exclude' and the reason(s) for exclusion noted. In cases where a clear decision on any of the inclusion/exclusion criteria could not be reached based on the abstract or summary alone, the publication was retained for additional review (Section 2.3.3). In total, 347 studies were for full-text review.

2.3.3 Full text reviews, quality assessments, and data extraction

In the next step, the full texts of all 347 potentially relevant documents were obtained and screened against the same inclusion/exclusion criteria as the abstracts. Every study that did not meet all of the inclusion criteria was excluded and the reason for exclusion noted (see Flow diagrams in Appendix 6). The texts marked for inclusion were subsequently read closely for data extraction. The recorded data included information about the research design, participants, intervention and comparator treatment (where applicable) and outcome measures, as well as the study's findings and any possible explanations put forward by the authors, limitations, and bottom-line conclusions (see data extraction sheet in Appendix 9 for full details).

At this point, quality assessments were also conducted for all included studies using Gorard's Sieve, a tool for assessing the trustworthiness of research findings from intervention studies (Gorard, 2014). Reviewers considered the overall study design, number of participants[5], rate of attrition, outcome measures, implementation of the intervention, and any other factors which may threaten the validity of the research findings. A score between 0 and 4 was assigned to each criterion (see Appendix 10) and the study was given a total score equivalent to the lowest of the sub-scores.

As with prior screening steps, a sample of 10% of the studies was appraised by two reviewers, iteratively until the rate of agreement exceeded 90%, which demonstrated consistency in applying the assessment tool. The team's trustworthiness ratings of the individual studies can be found in Appendix 11. These served to inform the narrative synthesis of the results with studies which were deemed to be of higher quality being given more weight than studies with greater methodological shortcomings.

In updating the review by Fitzpatrick, Morris, Clark et al. (2019), the trustworthiness ratings were additionally used to identify and exclude lower quality studies. In their original review, Fitzpatrick and colleagues used a Data Extraction Form adapted from the EPPI-Centre (2007) to assess the 'weight of evidence' in the identified studies as well as their relevance to the goals of their review. The authors excluded any studies which did not receive a 'HIGH' score for both relevance and weight of evidence. In the current REA, the team chose to exclude all studies from the update of this particular seed review which received a rating of 2* or below based on Gorard's Sieve. We acknowledge that there are differences in the methods by which the trustworthiness assessments were made between the original seed review and our update. However, given the timeframe allotted to this REA, it was not possible to reassess the trustworthiness of the individual studies included in the seed reviews so that all ratings

[5] 10-25 cases per comparison was considered a Small sample; Medium between 26-50 and Large > 50.

conformed to a common scale. We assume that the trustworthiness assessments are broadly valid across the different instruments.

Gorard's sieve is, however, a very strict assessment tool and Fitzpatrick et al. (2019) also took into account the relevance of the individual studies to the aims of their review. In order to avoid discarding studies which may have some design flaws, but which would nonetheless contribute significantly to answering the Review Questions of the current REA, the team decided to discuss any highly relevant but methodologically less sound studies in a separate section of the report. To identify relevant studies, the team applied the same criteria as Fitzpatrick et al. (2019) — that is, the relevant section of the EPPI-Centre (2017) Data Extraction Form — to the newly identified studies which received a rating below the aforementioned 3* Gorard's Sieve rating cut-off. Once more, 10% of the studies were double screened to demonstrate consistency in applying the assessment tool.

The REA by Fitzpatrick et al. (2019) was the only one of the seed reviews to exclude studies based on their quality and to assess the studies' relevance in a unified, formalised way. Although the team recognises that applying this method across all of the seed reviews would have increased the rigour of the current work, it was not possible within the scope of this REA to conduct quality assessments of all studies included in the original seed reviews.

2.4 Synthesis of findings

The review team has been tasked with addressing a series of questions that represent enormous scope and room for interpretation. Review Question 1, for example, asks 'What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?'. The first part of this question implies an exploration into all possible approaches to teaching foreign languages that have been promulgated and tested. The second part refers to effectiveness. Measures of effectiveness used in individual studies are likely to vary considerably depending on the type of intervention, the outcomes of interest, the population of interest, and the nature of the comparator. Both parts of this question, therefore, invite the likelihood of substantial heterogeneity between studies. This heterogeneity may be in terms of participants, settings, interventions, comparators, outcome measures, and so on. Moreover, some of the review questions ask for information that is best addressed through qualitative approaches to the data. For example, Review Question 5 asks 'Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?'. In addressing this question, it is more likely that useful information will be generated by reporting the descriptive findings of process evaluations in the primary research, rather than relying on bottom line findings based on quantitative data. As a result of this expected heterogeneity across studies and the nature of the review questions, we will follow guidance proposed by Popay et al. (2006) for narrative synthesis of heterogenous literature.

In brief, this approach to a narrative synthesis led us to generate textual descriptions of the primary research (both pre-existing based on information in the seed reviews and newly included studies) and organize these thematically. Themes were informed by the nature of each review question, and also emerged inductively in the process of preparing the REA. Indicatively, themes included teaching approaches, language domains, populations, settings, study designs, and the nature of the results being reported (e.g. different outcomes, or different implementation factors). The weight of the evidence based on the trustworthiness appraisals in the seed reviews and conducted by the team on newly included research were reported and discussed. Key features of individual studies were tabulated for visual comparison on a theme by theme basis. Finally, where sufficient data are reported in the primary studies, we created a common rubric to help understand how findings across individual studies compare to each other by calculating effect sizes (Hedge's G)[6] with confidence intervals for comparison interventions. Furthermore, we report findings with a 95% confidence interval or greater.

In our narrative synthesis of research findings, in section 3.1 we discuss evidence which speaks to RQ1 and 3 together, as these two questions logically address similar dimensions of the same question on what approaches are effective and which variables influence this effectiveness.

> RQ1. What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?

[6] g = the difference between means divided by the pooled and weighted standard deviation

> RQ3. What practitioner skills or programme characteristics contribute to effective language learning among students?

In section 3.2 we discuss the evidence addressing the second primary review question:

> RQ2. What is the impact of learning a foreign language on students' wider academic outcomes?

Finally, in section 3.3 we present the findings which address RQs 4 and 5, as these two questions (like RQs 1 and 3) are logically related.

> RQ4: What is the impact of using a non-native language as the medium of instruction in academic subjects on students' academic outcomes?
> RQ5: Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?

3. Findings

3.1 Review Questions 1 and 3

RQ1: What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
RQ3: What practitioner skills or programme characteristics contribute to effective language learning among students?

In this section we address review questions 1 and 3. These two questions are discussed together as we interpret RQ3 as being a subsidiary question to RQ1. In addressing these questions, we present our updates of the seed reviews carried out by Fitzpatrick et al (2018) and Harris and Ó'Duibhir (2011) as indicated in the methodology (Section 2). We begin by our discussion of the updates of the Fitzpatrick et al (2018) review, followed by our update of Harris and Ó'Duibhir (2011). We use our findings from both of these seed review updates to provide overall key conclusions to RQs 1 and 3.

We first provide a background description of Fitzpatrick et al (2018) and its main findings. We then present the updated studies we found since Fitzpatrick et al (2018) was published. This discussion highlights findings which speak to the effectiveness of different approaches to language teaching, including the amount and distribution of instruction time, the role of technology, and effective approaches to the teaching of vocabulary, grammar, and the four skills. After a summary of Fitzpatrick et al (2018) and our update to it, we present the findings of the next seed review which addresses RQs 1 and 3, namely Harris and Ó'Duibhir (2011). We identify the studies we found in our update to Harris and Ó'Duibhir organised around the themes of corrective feedback, intensive language programmes, and the development of L2 literacy. We bring the findings of our updates to both Fitzpatrick et al (2018) and Harris and Ó'Duibhir (2011) in the conclusions to section 3.1

3.2 Updates to the Fitzpatrick et al (2018) seed review

Fitzpatrick et al's (2018) rapid evidence assessment was commissioned by the Welsh Government "with the purpose of informing the Welsh Government's planning and delivery of Welsh language provision for learners aged 3-16 years, as it undertakes reform of curriculum and assessment arrangements in Wales" (p.7). In the context of policy drivers emphasising the need for a new school curriculum, an increase in the number of Welsh speakers, and ensuring that all learners would be able to use Welsh when they leave school, it was hoped that the review assessing research on language teaching practices and interventions used in international and national contexts parallel or comparable to Wales would create the potential for a paradigm shift in approaches to language teaching and learning. They posed one research question:

What teaching approaches and methods are effective in developing young learners' second language competence, according to high quality empirical evidence?

As the authors state in their Introduction, while its relevance to Wales was explicit, the REA was intended to make a significant contribution to teaching policy and practice in all non-dominant target language contexts.

The criteria for including research in the REA were that it should be:

- directly or indirectly relevant to language learners aged 3-16 years;
- directly or indirectly relevant to the context of teaching Welsh in Wales (for example, it has relevance to the teaching of non-dominant target languages);
- focused on "approach" and/or "method"; research on theoretical models, or teaching techniques/activities are only included if they are relevant to an approach or method;
- addressing deliberate, within class, teaching of second languages that are human, written/spoken languages.

The REA took place between November 2017 and March 2018. The initial scale and scope of the review included 12 areas of relevance as shown below:

i) Effectiveness of approaches/methods when applied to the young language learner context;

ii) Immersion and CLIL (Content and language Integrated Learning);

iii) Assessment of learning;

iv) Quality and intensity of learners' exposure to language;

v) Age and cognitive development;

vi) Practitioner skills and training;

vii) The processes by which 'transactional competence' develops;

viii) Development of bi- and multi-literacy;

ix) Cognitive and social advantages of language learning and bilingualism;

x) Motivation and attitude;

xi) Role of technology in language learning;

xii) Individual learner differences.

However, while it was acknowledged that all these areas were important for the development of policy, it was realised that these they would generate hundreds of thousands of results and the search was scaled back to the first two areas. The omission of the other areas helps to explain why areas of obvious relevance such as the role of technology and practitioner skills and training are given limited coverage in the review.

Search criteria were that the publication should be post-2000 (2001 being the European Year of Languages, which saw a substantial increase in publications relevant to teaching young learners), should be within social sciences or arts and humanities, be an article, book chapter, article in press, review or book, be in English or Welsh and be peer reviewed. Electronic searchers of 'grey' literature were also conducted.

At this stage, following the screening of 5861 items, 309 items were left to be assessed for eligibility by data extraction. Of these, 106 were included in the synthesis of evidence.
In Phase 4, the final phase, items and key findings were clustered by weight of evidence as highly trustworthy, highly relevant to the context of the REA, and appropriate to the REA in terms of research design following the data extraction form adapted from the EPPI-Centre (2007).

The themes that emerged from the synthesis of evidence were grouped as follows:

- Vocabulary competence (27 items)
- Grammatical competence (11 items)
- Reading competence (12 items)
- Writing competence (21 items)

- Speaking and Listening competence (24 items)
- General language competence (20 items)

These themes were followed in our synthesis of their included studies (apart from those that were related to CLIL and bilingualism) and our updated studies. The numbers of updated studies that were grouped into each of Fitzpatrick et al's (2018) themes are as follows:

- Vocabulary competence (9 items)
- Grammatical competence (5 items)
- Reading competence (3 items)
- Writing competence (2 items)
- Speaking and Listening competence (4 items)
- General language competence (0 items)

Table 3.1. Update studies found since Fitzpatrick et al (2018)

Study	Topic	Context	Sample	Findings (including effect sizes, where given)	Trust-worthiness rating in this REA
Chan (2018)	Comparing the effects of Processing Instruction, Traditional Instruction, and Implicit Instruction on the acquisition of the English simple past	Hong Kong primary schools	66 7-8 year-old pupils	On most post-intervention tests, the Processing Instruction group outperformed the Traditional Instruction and Implicit Instruction groups and showed the greatest gains. The range of the effect sizes between groups were *d*=0.06 to *d*=0.95.	3*
Chen, Liu & Todd (2018)	The effects of captioning on EFL learners' spoken vocabulary acquisition	Junior high school in Taiwan	118 8th grade students	Watching videos with captions led to a greater increase in form recognition (*d*=0.43) and form-meaning mapping (*d*=0.6). Participants with higher levels of linguistic competence performed better on form recognition (*d*=0.3). The presence of the captions assisted form-meaning mapping even among lower level learners (*d*=1.61).	3*
Gürkan (2019)	The effect of annotation use on vocabulary recall and retention levels among EFL students	Turkish elementary school	122 10th grade	Results indicated that the group who used multimedia annotations recalled and retained more lexical items than the other two groups, who read paper-based annotations of paper-based reading material; had had unannotated paper-based reading material respectively (Insufficient information to calculate Hedge's G but we calculated the multivariate effect size of =.387†).	2*
Kasprowicz, Marsden & Sephton (2019)	The effects of distribution of practice effects on the learning of L2 verb morphology	Seven English primary schools	113 8-11 year-old pupils	Results showed minimal group-level gains, yet there was substantial within-group variation in performance at post-tests. Individual differences in terms of accuracy of practice during training and language analytic ability were significantly associated (medium size effects) with higher post-test and delayed post-test performance under shorter (3.5 days) practice spacing	1*
Kasprowicz & Marsden (2018)	A comparison of two types of input-based practice for learning L2 German definite article case-marking cues	Three UK primary schools	138 9-11 year old pupils	Results indicated that both the form-meaning and the form-noticing interventions had large positive effects on case-marking of the accusative definite article in L2 German (*der/den*), and that the gains made were sustained nine weeks post intervention (*d*=0.01 to *d*=3.6). The control group made no gains.	3*
Meurers, De Kuthy, Nuxoll, Rudzewitz & Ziai (2019)	The effects of scaffolded feedback on the acquisition of L2 grammatical constructions	German high schools	205 7th Grade students (mean age 13.09 years)	On post-tests of conditionals, comparatives, and relative clauses, both groups had improved, with students in the *FeedBook* scaffolded feedback group significantly outperforming the students in the control group with a medium-size effect (*d*=0.56).	4*

Owen, Razali, Samad & Noordin (2019)	The effects of selected Communicative Language Teaching activities (Information gap; Language games) on EFL speaking performance	Libyan secondary school	124 1st year students	Analyses of scores on a pre- and post-test of 'oral speaking' found no difference between groups at pre-test, and that only the scores of Groups 3 (Information Gap) and 4 (Control) were significantly different at post-test (d=0.069).	2*
Padial-Ruz, García-Molina & Puga-González (2019)	Effectiveness of a motor intervention program on motivation and learning of English vocabulary	Honduran pre-school	88 4-7 year old pupils	Descriptive statistics indicated that the Gesture + Motor Activity group scored highest on the post-test, followed by the Control group and then the Gesture group respectively. No inferential analyses were conducted (thus no effect sizes were reported and insufficient information to calculate.).	2*
Park, Isaacs & Woodfield (2018)	Comparing extensive and intensive reading on EFL vocabulary development	South Korean secondary school	72 15-16 year-old students	Overall, students benefited significantly more from ER than IR in terms of knowledge of the meanings and uses of target words (η_p^2=0.08). There was a main effect of proficiency (η_p^2=0.28): advanced and intermediate level learners benefited more from ER, while low level learners benefited more from IR.	3*
Pujadas & Muñoz (2019)	The potential of extensive TV viewing for L2 vocabulary learning	Spain/Catalan high school	106 13-14 year- old students (8th grade)	All participants learnt vocabulary from extensive exposure to audio-visual input. There was a significant effect of group (ω^2=.199), with the captions-focused group being the most successful, followed by the subtitles-focused group, the subtitles non-focused group and finally the captions non-focused group. Proficiency level was significantly related to vocabulary gains in both form and meaning recall: more advanced learners obtained higher gains.	3*
Rostamian, Fazilatfar & Jabbari (2018)	The effect of planning time on cognitive processes, monitoring behaviour, and quality of L2 writing	Iranian private language school	60 intermediate students	None of the conditions successfully enhanced all of the quantitative measures simultaneously, yet there was a positive effect of on-line planning on accuracy (g=1.19 to g=1.88) and a positive effect of pre-planning on fluency (g=0.8 to g=1.14) and syntactic complexity (g=0.07 to g=0.95)†.	2*
Suárez & Gesa (2019)	The roles of proficiency and aptitude in learning L2 vocabulary from sustained captioned video viewing	Spanish/Catalan high school	57 Grade 10 students	A main effect for proficiency was observed on the learning scores for target words' (TW) forms and meanings: the higher a student's vocabulary size and listening score, the greater the gains in TW form knowledge when exposed to captioned video viewing. Language aptitude was only a significant factor for TW meanings. No effect sizes reported and insufficient information available to calculate effect sizes.	3*
Teng (2019a)	A comparison of *text structure* and self-regulated strategy *instruction* for L2 English writing	Hong Kong elementary schools	135 6th grade students	The Text Structure Instruction (TSI) and Self-Regulated Strategy Instruction (SRSI) groups outperformed the controls on essay writing (d=0.69 and d=0.83 respectively) and writing summaries (d=0.80 and d=0.65). Regression revealed that SRSI predicted better writing	4*

				quality, whereas TSI predicted more main ideas included in written summaries.	
Teng (2019b)	Effects of video caption type and word exposure frequency on incidental learning of L2 vocabulary	Hong Kong primary schools	257 Grade 6 (11-12 years old) pupils	The group viewing the full captioning video scored significantly higher than the keyword captioning group and the no-captioning group. The combination of full captioning and three encounters was most effective for incidental learning of lexical items. Effect sizes (η^2) ranged from 0.7 to 0.9.	3*
Teng (2019c)	Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos	Hong Kong primary schools	182 pupils, mean age 11.47 years.	Fully captioned videos were more effective for high proficiency learners' video comprehension, including global and detailed comprehension, but there was no significant difference between fully captioned videos and keyword captioning videos for learners with low proficiency. Both captioned formats produced better results than non-captioned videos. Both global and detailed comprehension were greater when videos were watched with captions twice and when words were shown more frequently in the keyword condition. No effect sizes reported and insufficient information to calculate.	3*
Van de Guchte, Rijlaarsdam, Braaksma & Bimmel (2017)	The effects of pre-task planning for video observations on L2 oral task performance	Dutch high school	48 students, mean age 14.2 years	The Focus on Language (FoL) group outperformed the Focus on Content (FoC) group on both attempted (*d*=1.3) and accurate (*d*=1.08) use of the target structure at initial post-test but this difference disappeared in the delayed post-test. Conversely, the FonC group outperformed the FonL group on the amount of coordination in the initial post-test (*d*=0.63) and on the subordination measure (*d*=0.57) in the delayed post-test.	3*
Van de Ven, Segers, and Verhoeven (2019)	L2 vocabulary learning through phonological specificity training	Dutch secondary school	86 students, 11-13 years old	Phonological specificity training led to increased word learning (measured via translation) compared to controls. Picture selection led to increased learning only for students with larger vocabulary sizes. Variances were reported but insufficient information to calculate effect sizes.	3*
Vyn, Wesely & Neubauer (2019)	The effects of foreign language instructional practices on student proficiency development	13 middle and high schools in United States	2,179 students	There was a largely positive effect for target language usage, which was most pronounced at beginner levels. However, teachers' reported use of explicit grammar instruction showed mixed results, beginning with a negative effect in Level I and moving toward a positive effect by Level IV. No effect sizes reported and insufficient information to calculate.	3*
Winasih, Cahyono & Prayogo (2019)	Effects of project-based learning using e-posters on EFL learners' speaking ability	Indonesian vocational secondary school	61 students	There was a significant difference between groups at post-test, with the experimental group (problem-solving group work using e-posters + group oral presentations) outperforming the controls (group text writing + group oral presentations) on an L2 speaking measure scored on content, vocabulary,	2*

				pronunciation, accuracy, and fluency (*g*=0.53†).	
Wang, Hwang, Li, Chen & Manabe (2019)	The effects on EFL learning of an integration of kinaesthetic technology and collaborative learning into total physical response	Chinese secondary school	79 11th grade (16-17 year old) students	The learners who used the Collaborative Kinaesthetic English Learning system outperformed the other two groups on a vocabulary test (*g*=0.51 and *g*=4.18)† but not on a comprehension test. No significant difference in learning of verbs, nouns or adjectives was identified.	2*

†Author (of this REA) calculated effect size

3.2.1 Research evidence on effective teaching of vocabulary, grammar and the four skills

We have followed the thematic categories of Fitzpatrick et al.'s (2018) assessment and have updated their included studies with more recent studies which meet our own criteria for inclusion.

The majority of the studies reported below were rated 3* and 4* by the REA team and are relatively small scale, of short duration, and focus on highly specific aspects of language teaching, in many respects mirroring the emphasis in the national curriculum for both KS2 and KS3 on specific content and skills rather than any particular approach. The 2* papers commented on are similar in focus and scope, yet outcome measures are largely self-developed by the researcher-authors, limiting their capacity for comparison of findings across papers, and randomisation is for the most part achieved only insofar as allocating intact classes to conditions. Some 2* studies do not use inferential statistics; of those which do, many do not report effect sizes.

3.2.1.1 Vocabulary teaching and learning

The main issue in most of the studies of vocabulary teaching and learning is whether it is more effective to rely on learners to acquire words and their meanings implicitly or incidentally from context or input alone, or whether there should be direct instruction, explicit practice and production of target words. Fitzpatrick et al. (2018) report several studies which compare outcomes of studies focusing on whether teaching vocabulary is more effectively done in context, known as 'Focus on Form' (FonF), through a communicative activity or reading text, through directly teaching lists of items ('Focus on Forms') or through translation.

For example, Laufer (2006) found that 16-year-old L1 Hebrew-speaking learners performed better (72% success) on target item recall when they had studied and practised items from a word list than when they had read a text containing the words, supported by dictionary look-up (47%). A follow-up study (Laufer & Girsai, 2008), differentiated not only between FonFs and FonF, but also between contrastive (L1-L2) and non-contrastive FonFs instruction. They found that 15-16 year old L1 Hebrew learners of English (n=75) receiving contrastive instruction in the form of L2>L1 and L1>L2 translation tasks, out-performed learners in the FonF and the non-contrastive FonFs conditions. It is worth remembering that the context for Laufer's work is her opposition to the 'focus on meaning' approach which relies on extensive comprehensible input for vocabulary acquisition with little direct instruction.

At primary school level, Shintani's (2013) subtle and complex study also compared the teaching processes and the effectiveness of FonF and FonFs approaches to teaching but with 6-year-old beginner learners' productive vocabulary knowledge during a nine-week intervention. She used a task-based teaching activity (each task involved learners carrying out instructions given by the teacher) to represent FonF, and a PPP approach (Presentation = repetition of target; Practice = drill; Production = in a game) represented FonFs. Target items were 24 nouns for which there was no significant difference in uptake between the two groups, and 12 adjectives, for which the FonF group out-performed the FonFs group.

Although both types of instruction were effective for the acquisition of nouns, the FonF instruction was found to be more effective for the acquisition of adjectives. Only the FonF learners developed the knowledge needed to use the adjectives in free production. Shintani proposes that differences between the process features of the FonF and FonFs instruction offer an explanation for this difference in learning outcomes since only the FonF instruction was characterised by contextualized input, the occurrence of negotiation of meaning, and student-initiated production.

As Alcón's (2007) study also confirms through the analysis of audio recordings from a year of English language classes, along with learner diaries from 14-15 year old Spanish/Catalan L1 participants (n=12), the effectiveness of vocabulary teaching which takes a FonF approach may depend on the timing and nature of teacher-led interventions. In her study, she identified instances of 'pre-emptive' and 'reactive' focus on form by the teacher and found that pre-emptive FonF led to ‘noticing’ items (as reported in learner diaries), and there was a positive correlation between noticing and post-test production of items.

An updated study which offers a different perspective by Van de Ven, Segers, and Verhoeven (2019) highlights the importance of phonological training in vocabulary acquisition in the L2. They report a short (15 minute) intervention involving 86 secondary school learners in the Netherlands, aged 11–13, all native speakers of Dutch. The participants were randomly assigned to one of three vocabulary training groups in which they heard unknown English words, such as ‘maze’ or ‘dice’. In the intervention group, their attention was focussed on both the sound and the meaning (selecting pictures with minimal sound differences), while in the control and other intervention group, the participants’ attention was focused on meaning, not sounds. None of the groups wrote the words down during the training; they only heard the words. The authors found that directly after the intervention students exposed to the phonological specificity training had learned more new English words compared to those in the active control condition. In the picture-selection condition, participants with relatively large vocabulary sizes also learned more new English words compared to participants in the active control condition. They conclude that learners benefit from a mixed L2 vocabulary training method that combines meaning-focused and form-focused elements.

Regarding the role of input in vocabulary acquisition, Fitzpatrick et al. (2018) also offer some example studies. Shintani (2011) (discussed at length below in Harris and Ó Duibhir Section) found no significant difference in performance on three of four post-task vocabulary tests when she compared 6-8 year-olds’ input-based instruction (listen-and-do card selection tasks, n=13) with production-based instruction (matched tasks but with cued production, n=11). However, in the same context, she (Shintani, 2012) found that when teacher input was modified and learners’ voluntary production of target items increased by repeating the input-based task nine times over five weeks, learners’ negotiation of input pushed vocabulary gains from an input-only task. A similar finding was reported by Luan and Sappathy’s study of 10-11 year old L1 Malay learners of English (2011). Hennebry, Rogers, Macaro and Murphy’s study of 262 14-year old learners of French (2017) found that vocabulary instruction (including in L1) after a listening activity led to more effective recall of vocabulary than a listening only condition.

The facilitating value of switching strategically into the L1 for vocabulary instruction has also been reported in several studies found in Fitzpatrick et al. (2018). Lee and Macaro (2013) investigated use of L1 in vocabulary instruction for recall and recognition of target items by 12-year old Korean learners of English. The teachers switched into Korean to give information about new lexical items for 223 learners, while 220 received English-only instruction. For acquisition and retention, the code-switching group significantly outperformed the English-only group in both recall and recognition. Camo and Ballester (2015) also report the facilitative value of using the L1 with 10-11 year-old learners of English learning 20 target items. Their experiment and control groups listened to a story, and as a target word appeared, the experiment group were shown and heard the word in both L2 and L1 (Catalan), whereas the control group were only exposed to the L2 item. While both groups performed similarly in immediate post-test picture-matching tasks, in the delayed tests, using L1 translation proved to have a statistically significant positive effect on young learners' long-term vocabulary retention.

A number of included studies focus on vocabulary learning through a task-based approach, often involving games. Huang, Willson and Eslami (2012) provide a meta-analysis of 12 studies including six with high school learners. They found that vocabulary gains were greatest where the involvement load of the task was high, where a combination of output tasks were undertaken, and where more time was spent on task. Involvement load is a motivational-cognitive construct; a task requiring learners to need, search for and evaluate the meaning of a word is interpreted as having high involvement load. Task-based learning (TBL) and digital game-based learning (DGBL) tend to produce high involvement load. In another meta-analysis which included seven studies involving young learners, Chen, Tseng and Hsaio (2018) found, perhaps not surprisingly, that the greater the fun and adventure-challenge component in a game, the more likely it was to be effective in terms of vocabulary uptake.

3.2.1.2 Using video, television and films with or without captions/subtitles for vocabulary acquisition

While there has been substantial research into the value of visual media in both formal and informal situations for decades, most of this research has focused on learners in higher education or post-formal education adults (e.g., Vanderplank (2010) and Yeldham (2018) for summaries). A meta-analysis by Montero Perez, Van Den Noortgate and Desmet (2013) confirmed the value of watching video material (TV programmes and films) with captions (same-language subtitles intended for the deaf and hard-of-hearing) for vocabulary acquisition and listening comprehension but contained no examples of studies among school-aged children.

Fitzpatrick et al. (2018) include one study on vocabulary learning through television programmes by Williams and Thomas (2017) which assessed uptake of Welsh vocabulary by 4-5 year-old English speakers in four 20-minute interventions applied three times weekly for 6 weeks. The interventions were: i) watching 15 Welsh TV programmes; ii) as i, but with teacher interaction; iii) "storytime", where the same stories as in the TV programmes were read aloud in Welsh; iv) as iii), but in English (control group). Post-tests on vocabulary from the programmes/stories found that the control group was outperformed by all other conditions. The highest vocabulary scores were seen for watching the TV programmes with teacher interaction, and there was no difference in performance between those who watched the TV programmes and those who listened to the same stories read by a teacher. The authors conclude that TV programmes, especially when viewed in interaction with a caregiver, can facilitate language uptake.

We have identified more recent studies which have reported positive findings in using captioned video material with school-aged learners for vocabulary acquisition. Chen, Liu and Todd (2018) compared the effect of watching videos with captions and without on vocabulary in a three month study in which 118 8th grade EFL learners in a junior high school in Taiwan, divided into a caption- and non-caption watching groups watched 10 videos in the series *Olivia*, an animated series for children. On pre- and post-tests of target vocabulary, the group that watched with captions achieved significantly higher scores (form recognition; Cohen's *d* =0.43; vocabulary acquisition, Cohen's *d* =0.60). Participants with higher levels of linguistic competence performed better (Cohen's *d* =0.30). A key finding was that the presence of the captions assisted aural word recognition; even lower level learners were able to perform "phonological recoding", transferring the visual input into its phonological code, and were able to transfer the benefits of visually presented textual input (captions) to oral tasks such as the vocabulary tests used in this study.

Pujadas and Muñoz (2019) report a year-long study in which 13-14 year old Spanish-Catalan bilinguals at low and intermediate (pre-A to A2 levels on the CEFR scale) levels watched twenty-four 20 minute-long episodes of Fresh of the Boat (an American series) in groups either with captions or subtitles (i.e., L1 translations) and with or without instruction focussing on 120 target words (40 per term, 5 per episode). Independently of the experimental condition, all participants learnt vocabulary from extensive exposure to audio-visual input. The groups who had been taught the target items beforehand performed better than the other groups who received no prior teaching groups; the captions-focused group was the most successful, followed by the subtitles-focused group, then the subtitles non-focused group and finally the captions non-focused group. Learners' proficiency level was significantly related to vocabulary gains in both form and meaning recall, with more advanced learners obtaining higher gains. The study provides valuable evidence that explicit instruction and extensive viewing is possible and effective, and that a small amount of teaching (instruction consisting of simple 5 minute-activities), directing learners' attention to target vocabulary may bring about significant improvement – especially on form recall.

Teng (2019b) investigated the effects of various captioning conditions (i.e. full captioning, keyword captioning, and no captions), the number of word encounters (one and three), and the combinations of these two variables on incidental learning of new words while viewing a video. Six possible conditions were explored with 145 target lexical items, involving 257 primary school students learning English as a second language. A post-test, measuring the recognition of word form/meaning and recall of word meaning, was administered immediately after participants viewed the video. The group viewing the full captioning video scored significantly higher than the keyword captioning group and the no-captioning group. The combination of full captioning and three encounters was most effective for incidental learning of lexical items (η^2 value showed effects (ranging from 0.7 to 0.9)).

Suárez and Gesa (2019) explored the effects of individual differences such as aptitude, listening skills and vocabulary size on extensive vocabulary learning over a term for 57 Grade10 EFL learners watching captioned video materials. On a weekly basis, all learners were pre-taught a set of target words (TWs); half of them (the experimental group) were additionally shown captioned episodes from a TV series, *I Love Lucy*, (mean length 24 m 30 s) containing the TWs. All learners were pre- and post-tested on the TW forms and meanings. Results revealed significant differences between experimental and control groups in the learning of TWs. A main effect for proficiency was observed on the learning scores for both TW forms and meanings: the higher a student's vocabulary size and listening score, the greater the gains in TW form knowledge when exposed to captioned video viewing. However, language aptitude was only a significant factor for TW meanings.

3.2.1.3 Intensive or extensive approaches to vocabulary development

We have also found a high-quality study by Park, Isaacs and Woodfield (2018) which compared the impact of Extensive Reading (ER) and Intensive Reading (IR) approaches by testing words actually contained within the reading texts and examining learners at a range of proficiency levels (advanced, intermediate, low). Seventy-two Korean secondary students aged 15-16 received either ER or IR teaching 2 hours once a week plus follow-up homework over a 12-week timespan, with pre- and post-performance differences examined by proficiency level. The experimental (ER) group received lessons based on a class library (a full set of 50 graded readers) for students to borrow. Each learner devised an individualized reading list of books and read about one per week, in class and as homework. Their reading logs showed they spent an average time of 164 minutes reading outside of class each week (with large differences in time spent).

The IR lessons involved analysing and translating four short (700-800 words) texts once a week for two hours. The teacher provided explicit instruction of reading strategies, grammar points and vocabulary relating to the texts. Participants were required to read four new reading texts per week and to complete further exercises. The IR group read an average of 192 minutes outside of class during the treatment and work was checked through quizzes at the start of each following class.

For the ER group, the entire text of each grade reader was scanned to enable tailored pre- and post-tests. Results showed that students benefited significantly more form the ER than from the IR treatment in terms of their knowledge of the meanings and uses of target words. Advanced and intermediate level learners benefited more from ER, while low level learners benefited more from IR. The authors suggest that teachers should carefully consider their learners' proficiency level when selecting a reading approach, in order to optimize learners' vocabulary development.

3.2.1.4 Integrating imagery, gesture and movement, and songs

The evidence provided by Fitzpatrick et al. (2018) on integrating imagery, gesture, movements and songs into effective vocabulary learning is limited and mixed. There are two studies which present some rare empirical evidence supporting the so-called 'Keyword' method of learning vocabulary, through attending to phonetic or orthographic features of the target item and linking these to a familiar 'keyword', usually in the L1. Research by Dolean (2014) with 101 Romanian leaners of Italian at primary and secondary level found that presenting learners with the image of the target word, alongside a keyword image, led to significantly better L2>L1 translation performance in an immediate post-test and in delayed post-tests (a significant medium sized main effect for treatment (F(1,99)¼18.33,p<.001, omega squared=.10), and a follow up study with 24 7-8 year olds and 21 13-14 year olds found a significant positive effect of keyword presentation, including in a delayed post-test (F(2,86)¼193.02,p<.001,omega squared=.49). A further study by Dolean and Dolghi (2016) found a Keyword-instructed group of 6-7 year-old Romanian learners of English (n=34) significantly outperformed a Total Physical Response-instructed group, with a large effect size on 30 imageable items [*F*(2,66) = 58.11 , *p* < .001, *ηp*
2 = .638].

In our update of studies, we found some 2*-rated studies which also focused on imagery and lexical acquisition. For example, Gürkan (2019) compared the vocabulary learning outcomes of Turkish elementary school pupils in three conditions: (1) a group who used a mobile app in with in-app, multimedia annotations (e.g., images; videos) of target lexical items in online reading material; (2) a group who read paper-based annotations of paper-based reading material; and (3) a group who had unannotated paper-based reading material. A researcher-developed vocabulary measure was administered at immediate and delayed post-test, with results indicating that the group who used the multimedia annotations recalled and retained more lexical items than the other two groups.

On the value of gesture and movement reported by Fitzpatrick et al. (2018), Porter's (2016) study found an initial significant advantage to gesture elaboration while teaching formulaic utterances to 4-7 year olds (n=40), but also that recall dropped considerably (though remained higher than control group scores) at a 2-week delayed post-test. In a more complex study, Mavilidi, Okely, Chandler, Cliff and Paas (2015) compare four learning conditions for teaching 14 Italian words to 111 children (mean age 4.9) in Australia. The conditions were simultaneous to visual and oral word presentation, and were: integrated physical exercise (children enact actions); non-integrated physical exercise (unrelated to item); gesture (gestures to act words while seated); conventional (repeat words while seated). Free- and cued-recall test scores were low for all conditions. The integrated group performed significantly better than other groups for free recall, but still their average score was below three out of 14 words recalled. In cued recall, no significant difference was found between the two physical exercise groups, but they both performed significantly better than the other conditions, and the gesture condition produced higher scores than the conventional condition.

In terms of updated 2*-rated studies on movement/gesture and lexis, Wang, Hwang, Li, Chen and Manabe (2019) looked at the effects of a Collaborative Kinaesthetic English Learning (CKEL) system on vocabulary learning among eleventh-graders (16-17 years of age), presumably in China (as this is where the first author is listed as working) but this is not explicitly stated. The CKEL system was developed using Microsoft Kinect to integrate kinaesthetic technology (e.g., as in the Nintendo Wii) and collaborative learning into the Total Physical Response (TPR) L2 pedagogy. 48 target lexical items were presented in a series of games and activities on the CKEL or via more traditional methods such as through watching videos, with outcomes measured via a vocabulary test and a multiple-choice comprehension test. The learners who used the CKEL system outperformed those who did not on the vocabulary test but not on a comprehension test (no effect sizes reported). The authors explain this finding as follows: "to score high marks in sentence learning, one has to master a larger amount of vocabularies and grammatical rules. Hence, more attention should be directed towards the improvement of this system design so as to help learners to study sentences better" (2019, p. 771). Similarly, Padial-Ruz, García-Molina and Puga-González (2019) analysed the effect of a five-week physical activity and gestures intervention on the L2 English vocabulary learning and motivation of pre-schoolers (4-7 years old) in Honduras. The sample were absolute beginners and were taught 22 target lexical items related to the human body. The control group teacher used flashcards and spoke each item in both languages (e.g., brazo/arm). In experimental group 1 (Gesture), the teacher did the same but also pointed to the body part in question, encouraging the children to do the same. In experimental group 2 (Gesture + Motor Activity), the teacher read aloud the item in both languages; pointed to the body part, encouraging the children to do the same; and led the children in a physical movement making use of the body items (e.g., swinging the arms). Descriptive statistics indicated that the Gesture + Motor Activity group scored highest on the post-test, followed by the Control group and then the Gesture group respectively. No inferential analyses were conducted.

Fitzpatrick et al (2018) give an almost passing mention to the value of songs in language learning in schools, briefly including one study by Davis (2017); we consider that this issue deserves greater attention., Davis (2017), in a critical review of the evidence for the uses of songs in language learning and teaching for young learners (ages 3-12), highlights the fact that teaching materials for songs rarely draw on support from empirical research, and that classroom-based studies are greatly lacking. Drawing on a Google search which identified 200 potential items, he reports the evidence from nine studies that met inclusion criteria regarding the implementation of songs in the classroom and the assessment measures used. These nine studies included students with five different L1s representing multiple language families, and covered young learners from ages 3 to 12 in ESL and EFL environments in eight different countries. Davis concludes that while the evidence is limited by the scarcity of empirical research and small sample sizes in young learner classrooms, the overall pattern of findings suggests that songs have pedagogical value and may promote both receptive and productive vocabulary acquisition, increase motivation, and improve pronunciation, communicative abilities, and literacy.

Our conclusions are broadly in line with those of Fitzpatrick et al.'s (2018) on effective approaches to teaching vocabulary:

Focusing attention and intentional learning activity on form and meaning of individual vocabulary items enhances vocabulary uptake, but this must be strategically applied.
While input-only instruction is effective in terms of (limited) vocabulary uptake, learning gains are greatly enhanced when input is supplemented with some pre-teaching and further interaction.
Vocabulary learning is facilitated by tasks with high involvement load.
The use of well-selected and graded L2 video clips and TV programmes, especially with L2 captions and some guidance from teachers, is an effective means of learning new vocabulary.
Integrating creative imagery can boost vocabulary uptake as may songs; integrating gesture and movement yields more modest gains.
Timing and variety of mini-interventions in learning has a significant impact on effective learning, regardless of approach/method.
Teacher experience and competence is a key variable in successful vocabulary teaching and learning.
Teachers should carefully consider learners' proficiency level when adopting extensive or intensive reading approaches to vocabulary development.

3.2.2 *Effective approaches to teaching grammar*

Below we provide a summary of studies which focus on explicit instruction, oracy and literacy, distribution of practice and language analytic ability, and processing instruction.

3.2.2.1 Oracy and literacy

Reported by Fitzpatrick et al. (2018) and of particular relevance to the context of the present REA is Graham, Courtney, Marinis and Tonkyn's (2017) study of the relative merits of an oracy-based and a literacy-based approach to teaching primary school children aged 9-10 (n=252, though actual numbers varied at the three different test points). Learners completed a sentence repetition (SR) task and a photo description (PD) task, making small but statistically significant progress in both grammatical and lexical knowledge between test points (SR small to medium effect sizes, *d*= .34 to .79; PD, *d*= .27 to .70). They found that there was little difference in outcome between the two. Learners with lower L1 (English) literacy scores, however, were benefitted slightly more from a literacy approach to the teaching of French. The study found that the teacher's level of training and the number of hours of instruction were far more important variables than type of instruction. Pupils with a teacher with degree-level French made significant progress in grammatical competence at all test points, whereas those whose teacher had GCSE-level French or below failed to progress between school years 5 and 6. Those who received 60 minutes or more of instruction per week achieved test scores in year 5 that other students, who received less instruction per week, began to achieve only much later, in year 7. The authors conclude that type of instruction is not a decisive factor in children's grammatical development in the L2 between primary and secondary school.

In a rare study investigating whether prosodic features in continuous speech may help to reinforce the grammatical functions of different word types (the so-called bootstrapping hypothesis), Campfield and Murphy (2017) found that providing eight-year-old learners of English with input rich in prosodically-marked features led to better results on a GJT testing understanding of English word order.

3.2.2.2 Explicit Instruction

We follow Fitzpatrick et al.'s (2018) use of the term "explicit instruction" to mean any kind of instruction in which learners come to an explicit awareness of target language features. In deductive instruction, these features are brought directly to learners' attention by the instructor; in inductive instruction, they are led to discover the target features for themselves, and these are later confirmed by the instructor.

On the whole, the evidence base reported by Fitzpatrick et al. (2018) supports explicit instruction of grammatical structures over learners becoming aware of such structures implicitly. Below are several studies which met their standard for inclusion. For example, Hanan (2015) found that explicit instruction with either a focus on form-meaning correspondence, or on form only, was effective for learners of German aged 9-11 in the input-poor environment of three English primary school classes, who made substantial gains on written, oral, and metalinguistic tasks. She also reports that a sub-group within each class accounted for most of the group level gains. This advantage of explicit instruction was supported by Tode (2007), who found that explicit instruction for 12-13 year-old Japanese learners of English led to immediate performance gains compared to learners who followed an implicit learning approach, but that these gains did not persist to a delayed post-test.

As the findings of Lichtman's (2013) study with children aged 8-17 learning Spanish in the United States indicated, explicit instruction was effective for tasks that tested explicit knowledge of grammar, while an implicit approach led to stronger performance in tasks believed to tap implicit knowledge, such as a story rewriting task. In similar vein, a small scale and short-term study by Toth and Guijarro-Fuentes (2013) provided evidence that explicit instruction leads to improvement in tests that targeted implicit knowledge in the use of 'se' of 3rd year Spanish learners aged 15-17 in an American high school. On another scale and of particular relevance to the present context, Tammenga-Helmantel, Arends, and Canrinus' (2014) study of 981 Dutch children in 42 secondary schools aged 12-15 learning English, German or Spanish found that any kind of exposure to a target form, whether explicit, implicit, or incidental, led to gains on grammaticality judgement tests (GJTs) and the correct use of the target item in writing tests. However, the GJ-gain scores and the scores in the performance test gave equivocal outcomes concerning the position of explicit instruction compared to the incidental and implicit methods and for the specific language(s) to which these findings apply. They found a weak preference for explicit instruction in two contexts only: GJT scores for learners of English, and writing test scores for learners of German. When Tammenga-Helmantel, Bazhutkina, Steringa, Hummel and Suhre (2016) treated deductive and inductive methods as variables with learners aged 15-18, they found that inductive instruction was slightly more effective for performance on a GJT, but not on a writing test.

In a comparative study of methods lasting four months, Ho and Binh (2014) report that that both traditional grammar translation method (GTM) teaching and communicative-style explicit instruction were found to increase grammatical knowledge in participants aged 12, although only the communicative-style instruction led to gains on an oral production task. Ho and Binh (2014) also found that an inductive teaching method led to gains in grammatical competence on both a grammar test and an oral production test.

In an updated study (overall rating 3*), Kasprowicz and Marsden (2018) report a matched-pair randomized experiment "to determine the effectiveness of explicit instruction for developing young learners' grammar within input-poor FL classrooms, in a context in which the curriculum demands grammar teaching" (2018, p. 888). Participants (n=138, aged 9 to 11) in 7 classrooms in 3 primary schools in England with beginners level German received instruction in two forms: explicit information followed by either task-essential (TE) practice in making form-meaning (FM) connections (referential activities such as ‘who is doing the photographing?’ in response to a picture and caption) or task-essential practice in spotting the form (F) (noticing activities), to test the usefulness of repeated, task-essential attention to forms in the input. Case-marking of the accusative definite article in L2 German (*der/den*) was the target feature. Learners from four classes were assigned to either the TE-FM (n=46) or TE-F (n=41) groups using matched pair randomization based on their composite score on two written pre-tests. The three remaining classes formed a non-active control group (n=52). During the experiment, the Control group continued their normal German lessons. The results of a battery of six tests of oral and written comprehension and production showed gains for the intervention groups at post-test and delayed post-test compared to no gains in a control group (pre-and post-tests only). These results indicated that both the form-meaning and the form-noticing interventions had large positive effects (Cohen’s d = 3.60) on the learners’ comprehension and production of the feature. Across all measures, no differences were found between the treatment groups at post- or delayed post-test. Crucially, no gains were observed in the control group, indicating no test effect. The authors suggest that practice in both noticing forms and making form-meaning connections can lead young learners to develop knowledge that is accessible under some time and communicative pressure and after a two-month delay.

In contrast to the above studies which tended to show positive effects of explicit grammar instruction, in our update of research, we found Vyn, Wesely and Neubauer’s (2019) investigation (rated 3* overall, see Table 6.11 Appendix 11) into the relationship between teachers’ practices and students’ proficiency development in a large cohort of secondary school students in the United States learning foreign languages (n=2,179) showed mixed results on assessment of year-long gains in listening, reading and writing. Twenty-six teachers of French and Spanish from 13 middle and high schools reported their target usage and explicit grammar instruction through a targeted web-based survey. Their findings showed a largely positive effect for target language usage, which was most pronounced at beginner levels. However, teachers’ reported use of explicit grammar instruction showed mixed results, beginning with a negative effect in Level I and moving toward a positive effect by

Level IV. Collectively, the findings suggest that best practices for foreign language teaching may vary according to the level of instruction.

In another updated study with a lower overall rating (only 1* overall), Kasprowicz, Marsden and Sephton (2019) investigated whether distribution of practice at different intervals and language analytic ability (the capacity to infer language rules and make linguistic generalizations and extrapolations) made a difference to the effectiveness of explicit, input-based grammar instruction for young learners of French (aged 8 to 11, n=113)) in seven English primary schools. Intact non-randomised classes were used. The group which received instruction at seven-day intervals included two mixed Year 5/6 classes (ages 9–11) and one Year 5 class (ages 9–10). The 3.5-day interval group included two Year 5 classes and one Year 4 class (ages 8–9). The control group included one mixed Year 4/5/6 class (ages 8–11) and one mixed Year 5/6 class. This group completed the tests only and reverted to their normal French lessons between pre- and post-tests. There was a degree of attrition in those taking the battery of post-tests. The treatment groups undertook identical tasks, both totalling 180 minutes but differing in the distribution of the sessions. Training for the 7-day and 3.5-day groups was delivered via a bespoke, digital, game-based application containing a series of mini-games, with each game teaching just one particular grammatical contrast expressed by one pair of inflections such as first person singular vs. plural present tense inflections. Training was completed on individual laptops with headphones. Their findings indicated minimal differences between longer (7-day) versus shorter (3.5-day) spacing of practice for learning a French verb inflection subsystem at either post-test or delayed post-test. While there were minimal group-level gains, there was substantial within-group variation in performance at post-tests. Individual differences in terms of accuracy of practice during training and language analytic ability were significantly associated with higher post-test performance under both forms of practice spacing. While this study scored highly in terms of ecological validity, its design was a major limitation.

3.2.2.3 Processing instruction

While Fitzpatrick et al. (2018) excluded studies on the effectiveness of Processing Instruction as an approach to teaching grammatical features, including Shintani's meta-analysis of 42 experimental studies in 33 articles (Shintani, 2014), a later publication by Chan (2018) met our criteria for inclusion. Shintani defines Processing instruction (PI) as follows: 'PI minimally consists of ‘structured input’ activities but can also involve explicit information (EI). Structured input consists of a set of exercises requiring learners to process the form-meaning mapping of a specific grammatical feature in the input and then demonstrate that they have done so’ (2014, 307).

Chan (2018) compared the Processing Instruction approach to teaching the English simple past tense with what she termed Traditional Instruction and Implicit Instruction in three roughly equal groups (n=66) of Primary 2 (7-8 years old) pupils in Hong Kong in a three-day intervention of 1½ hours’ instruction each day. The pre- and post-intervention tests involved reading, listening and ‘fill in the blanks’ writing. On most post-intervention tests, the PI group outperformed the other groups and showed the greatest gains. Overall, however, Chan rightly suggests that there is a role for each approach in teaching the forms and meanings of difficult and hard-to-grasp grammatical concepts for young learners in the initial stages of foreign language acquisition.

Hanan’s (2015) study of 9-11 year-old children learning a difficult German structure summarised above (Grammar: Explicit Instruction) is also an example of a teacher explicitly using PI successfully in terms of comparative gains.

There is a striking absence of included studies on the role of technology in supporting grammatical instruction in Fitzpatrick et al. (2018) given the number of high-quality journals which focus on technology and language learning. In updating studies, we found only one study which met our criteria: Meurers, De Kuthy, Nuxoll, Rudzewitz and Ziai (2019) carried out a randomized controlled trial in 7th grade classes (mean age 13.09 years) in four German high schools in which the authors test an intelligent tutoring system, a specially designed web-based workbook, *FeedBook*, compared to a printed workbook widely used in German schools. *FeedBook* provided immediate scaffolded feedback to students on form and meaning for various grammatical exercise types given for homework. 104 pupils took part in the intervention using *FeedBook* to scaffold homework on conditionals, comparatives, and relative clauses, and 101 pupils in the control group. On post-tests of the grammar constructions, both groups had improved, with students in the *FeedBook* scaffolded feedback group significantly outperforming the students in the control group, indicating that scaffolded written feedback on forms is an effective intervention method. Providing secondary school students with immediate scaffolded feedback on grammar while they work on their homework significantly improved their mastery of those grammar aspects.

In summary, the evidence from studies included by Fitzpatrick et al; (2018) and our updated studies for effective approaches to teaching grammar indicates that:

- While explicit instruction in grammatical features is effective, it is rarely more effective than other types of instruction in developing grammatical competence and should be tailored according to age and level of proficiency.
- Both inductive and deductive types of explicit instruction are effective, but inductive may be slightly more effective under certain conditions.
- Attention to prosodic features (e.g. rhyme, rhythm) in oral input can aid the development of grammatical competence.
- Both oracy and literacy approaches in primary school can be effective in developing grammatical competence.
- Individual developmental differences in ability to handle linguistic concepts should be allowed for in primary school level teaching.
- Teacher language competence, experience and number of hours' instruction are more influential factors than instruction type.
- Tutoring systems which provide immediate feedback on grammatical exercises can assist learners in achieving competence in grammatical structures.
- Processing instruction may be worth considering as a structured approach to presenting difficult grammatical concepts.

3.2.3 Effective approaches to teaching the four skills

3.2.3.1 Reading skills

Fitzpatrick et al. (2018) report research into effective approaches to teaching reading in a foreign language under lower-level processing skills, such as orthographic decoding, and higher-level processing skills, including strategy instruction and cognitive load factors. In addition to these studies, we have included research on intensive versus extensive reading practice. Three studies reported by Fitzpatrick et al. (2018) confirm the value of phonics instruction for reading words, correct pronunciation as well as fluent and accurate reading aloud, if not improvement in comprehension.

Lower level processing skills

In Takeda's (2002) study, which compared a treatment group of 12-13 year old Japanese learners of English which received six months of phonics instruction with a control group, phonics instruction was shown to be a significant factor in the treatment group's improvement on tests of pronunciation and reading aloud.

Similar results were found by Fonseca-Mora, Jara-Jiménez and Gómez-Domínguez (2015) in an 11-week intervention with 7-8 year-old beginner English language students in Spain. They compared a control group (in which the teacher used the syllabic and global word approach) with two experimental groups receiving phonological training, one with music support. The phonological training programme included phonics instruction, and also phonological awareness development, particularly of sounds which were not distinguished in the learners' L1. Students in both treatment groups performed significantly better than students in the control group on tests of naming upper and lowercase letters presented randomly to the students and identifying the initial sound of ten words read aloud. A non-significant trend was also reported to be found in the non-music phonological treatment group for the largest improvement on a test of reading a dialogue with accuracy, speed and fluency.

Porter (2014) also included systematic and explicit phonics instruction in her single cohort study of 9-11 year-olds learning French in England. Her 23-week study looked at four main elements: simultaneous oracy and literacy development, a focus on L2 sounds, phonics instruction, and L2 sound and print. The training also involved attention to differences between the L1 and the target language. Porter found statistically significant increases in both reading aloud and reading comprehension scores in week 21 of the intervention. As there was no control or comparison group in this study, only limited conclusions could be drawn.

Included in our update of relevant studies, Chen, Liu and Todd (2018) (reviewed above under Vocabulary) found that junior high school learners in Taiwan watching 10 videos in the series *Olivia* were able to perform "phonological recoding", transferring the visual input into its phonological code, and were able to transfer the benefits of visually presented textual input (captions) to oral tasks such as the vocabulary tests used in this study.

Fitzpatrick et al. (2018) report two studies on effective approaches to developing oral reading fluency which compared the use of tablets with a stylus or digital pen with paper-based, peer- and teacher-supported approaches. In Lan, Sung and Chang's (2009) study involving 9-10 year-old English learners in Taiwan, instruction included individual learning of phonics rules and vocabulary and reading a paragraph of a text and cooperative learning such as teaching each other the learned rules and vocabulary or putting paragraphs into the right order, as well as peer and teacher assessment. Oral reading fluency was tested before and after the 10-week treatment and there was no significant difference between groups, both of which made improvements. The raw scores suggested that in the control group the instruction mainly benefited high-level ability students, while the computer-assisted instruction benefited most students. Video analysis of behaviours during the intervention showed that the treatment group was more focussed on their activities, while the control group spent time chatting, walking around and playing, and were generally more teacher-dependent.

Chen, Tan and Lo (2016) also studied Taiwanese English learners, this time aged 13-14. In this 8-week study, the treatment group practised repeated reading with the support of a digital pen, while the control group practised with the support of a peer and/or the teacher. The main difference between the groups was that the digital pen allowed students in the treatment group to record their own readings and listen back to them, thereby encouraging self-assessment in comparison to the model. Both groups of students significantly improved on oral reading fluency tests between pre- and post-tests. However, the experimental group made more significant progress than the control group. Interview data with a sample of students from the treatment group pointed to the benefits of self-learning through the digital pen system in terms of learners' active engagement with and control of their learning.

Higher level processing skills

Six studies of reading strategy instruction are reported by Fitzpatrick et al. (2018), all finding that it can assist in developing comprehension skills (Harris, 2007; Macaro & Erler, 2008; Macaro & Mutton, 2009; Manoli, Papadopoulou, & Metallidou, 2016; Martínez & de Zarobe, 2017; Mistar, Zuhairi, & Yanti, 2016).

The studies cover groups ranging in age from 10-11 to 15-16 and in all cases, treatment groups outperformed control groups on measures of reading comprehension ability. The strategies chosen for instruction varied between studies, but many common ones such as predicting text content, using prior knowledge, skimming and scanning, and inferring/guessing were used in multiple studies. All but one (Macaro & Mutton, 2009) of the six studies of strategy instruction involved the teaching of more than one strategy simultaneously. All interventions were successful in improving students' comprehension skills, though Harris (2007) found that the very large number of strategies in her study caused some problems for the learners, including difficulties in remembering all the different strategies and difficulties in selecting the most appropriate strategy from the large repertoire. No such problems were reported in other studies involving multiple strategies, suggesting that explicitly teaching no more than six strategies at a time might be more appropriate.

Two studies investigated the effect on reading comprehension of manipulating cognitive load and glossing foreign language words, one using sophisticated technology, the other a novel approach to creating language learning materials for beginner learner while developing their inferring strategies. Türk and Erçetin (2014) investigated the use of multimedia glosses during reading comprehension tasks in a computer lab with high level Turkish students (B2 level) aged 15-16 in which they compared two conditions of learner control over the presentation of glosses: learner choice of text, graphics or both; and simultaneous presentation of text and graphics (with no learner choice). The study found that students in the simultaneous presentation group outperformed those who had a choice of glosses in reading comprehension tests. The authors conclude that this is explained by a reduction in cognitive load, implying that materials which are adapted in order to reduce the distribution of cognitive load may lead to better learning.

Macaro and Mutton (2009) carried out a novel intervention with Year 6 learners of French in which they compared reading comprehension development in two treatment groups using either adapted graded readers or an age-appropriate English story text with target language words embedded into the text in gradually increasing proportions. The English text group also received strategy training. Students in both interventions made significant advances in reading comprehension in comparison with a control group who received normal teaching provision rather than dedicated reading time with adapted material. The authors conclude the English text helped learners notice and acquire 'little words' (function words) which can be important for comprehension but that might be overlooked in L2-only texts by reducing the cognitive load on working memory.

From the evidence of above research studies included by Fitzpatrick et al. (2018) and updated for this REA, effective approaches to developing reading skills may be summarised as follows:

Effective approaches include explicit attention to both lower-level and higher-level processes since it cannot be assumed that either will be transferred from a student's L1 without instruction.

Phonological training can help beginner learners to process word forms, but not necessarily word meanings.

Technology-supported learning can contribute to the development of oral reading fluency by facilitating student-centred learning.

Watching video material supported by captions can assist learners in aural word recognition and "phonological recoding" of foreign language aural input.

Computer-assisted learning of reading may enhance focus on task, reduce teacher dependency, and benefit lower ability learners.

Explicit instruction in reading strategies and skills is effective in developing reading comprehension skills.

A range of strategies can be taught together, but not in excessive quantities.

Reading materials which are adapted to reduce cognitive load can help to support comprehension.

3.2.3.2 Effective approaches to teaching L2 writing skills

Fitzpatrick et al. (2018) include 21 studies in their synthesis which they group under the role of technology, non-traditional pedagogical approaches and strategies for developing writing skills. Most of the studies on the role of technology are qualitative or involve single groups of learners. In more recent literature, we found numerous studies in which student writing was supported or developed through technology but none met the criteria for inclusion being either poorly designed, single group or purely qualitative.

Typical of such studies is one included by Fitzpatrick et al. (2018) by Fidaoui, Bahous and Bacha (2010) in which they explore of the use of CALL (Computer-Assisted Language Learning) for English language writing with forty-eight 9 and 10 year-old Lebanese learners of

English over a 3 month period. The majority of participants felt the use of CALL was enjoyable (95.8 %) and motivational (64.6 %) for their writing development as they were better able to express their perspectives, gather and synthesise online information and develop “creative, neat, organised, error-free written products” (Fidaoui et al., 2010, p. 164). However, initial learner weaknesses in elementary computer and research skills prevented them from producing high-quality written work and the authors recommend careful planning of ICT-based writing work, and that students are properly monitored throughout.

Yunus, Nordin, Salehi, Embi, and Salehi’s (2013) survey of teachers’ use of CALL also noted the advantages of using ICT to stimulate student interest, develop their lexical knowledge and promote practical learning but also highlighted the difficulty of controlling the class, the ease of distraction and the tendency of pupils to write short-form responses to tasks. Teachers were also somewhat weak at handling these problems, and that planning to use computers in writing sessions was less than adequate.

Other studies included by Fitzpatrick et al. (2018) focus on more specific elements. For example, in a study involving 14-15 year-old English learners of German, Taylor, Lazarus and Cole (2005) investigated the use of drop-down menus to provide writing frames, which assisted students by providing options for part of a sentence they were attempting to formulate. Students were found to write at greater length, increase their accuracy and engage with tasks more enthusiastically. Progress was supported by PowerPoint presentations, in which grammar was addressed.

Weblogs, mobile devices and social media also figures in studies included and their findings indicate that while using such technological affordances may increase motivation, sense of achievement, engagement and enjoyment, those without the necessary skills or familiarity with the technology may feel marginalised and less enthusiastic. Sercu (2013) found that encouraging learners to contribute to an online blog led to a sense of achievement and successful collaboration, and increased motivation, though students who were weaker than their classmates at using technology felt marginalised and were less enthusiastic about participation. Hwang, Chen, Shadiev, Huang, and Chen (2014) found that using mobile devices increased motivation to learn writing skills in class, and subsequently encouraged achievement. While using mobiles for ‘situated writing’ about three familiar contexts (classroom, meal and playground) was challenging for the elementary students involved, those who were asked to use mobiles demonstrated a higher performance when asked to describe the environment and express their ideas than those who were not. In social media use, the findings of Buga, Capeneata, Chirasnel, and Popa’s (2014) study indicated that Facebook may be a useful tool for developing writing skills such as helping students experiment with learning methods and completing written homework tasks, when they had not previously done so.

Other approaches to writing instruction

Several studies are also reported by Fitzpatrick et al. (2018) which may be considered as ‘non-traditional’ approaches to writing skills instruction, namely, flipped learning, the process approach to writing and dialogue journal writing. Abdelrahman, Dewitt, Alias and Rahman (2017) looked at ‘flipped learning’, which involves giving students online materials prior to class, and using the lesson time to deepen understanding of these, for example through collaborative problem-solving activities. They found that writing proficiency and pupil engagement improved using this approach, particularly due to the interactive nature of the tasks set. The teacher was able to allocate more class time to help the learners, which was cited as an especially positive outcome. The authors note the possibility that such an approach might not suit every student and it is well known that variability in learners carrying out the out-of-class tasks may cause classroom issues. The impact of a process-based approach involving idea generation, multiple drafts, emphasising the reader, collaboration and creativity was evaluated by Ngo and Trinh (2011). Process writing helped to increase student writing performance and their enthusiasm for written work. The authors claim that focusing on content over correcting grammatical inaccuracies was particularly instrumental to motivation, in addition to prioritising a strong communicative message over error avoidance. Lastly, Ghahremani-Ghajar and Mirhosseini (2005) found that dialogue journal writing empowered students, fostered critical awareness and helped to develop the notion of ‘voice’.

Bartan (2017) explored the improvement of Turkish learners’ English language writing through short story reading, whereby using reading to provide a model for writing was found to help students develop their language, content, organisational structure and communicative achievement. Collaborative translation was also investigated as a potential means of instruction by Bruton (2007), who found that these tasks increased vocabulary development as students wrote and they also became able to identify different text formats they were exposed to.

Within the context of primary school level MFL teaching in England, where evidence suggests that an emphasis is generally put on oracy (Cable et al., 2010), teaching literacy and oracy together has also been suggested as an innovative means of instruction, and one which is feasible without one element necessarily having a negative impact on the other (Porter, 2014). Porter's study of 45 beginner learners of French, aged 9-11, in two primary schools identified first language (L1) reading age and verbal working memory as highly important factors in L2 oracy and literacy education. Porter concludes that all ranges of ability are capable of participating in L2 instruction in both oracy and literacy, and progress can be made by lower level pupils too.

In a 2* rated study from our update of Fitzpatrick et al. (2018), Rostamian, Fazilatfar, and Jabbari (2018) explored the effect of planning time on cognitive processes, monitoring behaviour, and L2 writing quality among junior high school students in Iran, who had intermediate proficiency in L2 English. All participants were required to write a narrative in English in response to a six-part picture story. The control group (Group 1) were required to write at least 200 words within a time limit of 15 minutes and were given no planning time (i.e. online planning with time pressure). Group 2 also had to start writing immediately, but were given no time limit and no word count to adhere to (i.e. online planning without time pressure). Group 3 had ten minutes' planning time, plus the time limit and word count (pre-planning with time pressure). Group 4 had ten minutes' planning time and no time limit or word count (pre-planning without time pressure). Each student was videotaped individually as they carried out the task, and these data were used as stimuli in subsequent stimulated recall interviews. The texts were analysed in terms of syntactic and lexical complexity, accuracy, and fluency (syllables transcribed per minute; token count divided by clause count). Findings revealed that none of the conditions successfully enhanced all of the quantitative measures simultaneously, yet there was a positive effect of on-line planning on accuracy and a positive effect of pre-planning on fluency and syntactic complexity.

There are number of studies which focus on teacher or learner strategies for developing writing skills effectively. In an updated study involving 135 6th grade students from 3 Hong Kong elementary school, Teng (2019a) placed students in one of three conditions: text structure instruction (TSI), self-regulated strategy instruction (SSRI) and a control group receiving 'traditional' instruction. TSI included instruction about text structure strategies while SRSI included self-regulation writing strategies, text and genre knowledge. The intervention lasted one month, after which time writing was assessed using measures such as writing quality and summarization of main ideas. The results indicated that the TSI and SRSI groups performed better on the tests than the traditional instruction group. Each novel technique had a different impact: SRSI on writing quality, TSI on the main ideas in the written summaries. A further linguistic and textual analysis of the two intervention groups writing showed high syntactic complexity, content organization and lexical variation in their compositions.

In another study reported by Fitzpatrick et al. (2018) comparing the effect of different approaches on written work, Gündüz and Ünal (2016), investigating the importance of multiple intelligences in 6th grade elementary school learners in Turkey, found that after 9 hours of classes on comparatives in English following either a grammar and rule-based approach or a content-based approach involving a very wide variety of activities to suit multiple intelligences, accommodating multiple intelligences in the classroom improved writing acquisition more effectively than the rather rigid, grammar-based approach. In addition, participants reacted more positively to this when compared to regular practices.

Other strategies taught to assist learner writing have included mind-maps, brainstorming and pre-writing planning, facilitated by using computers (Lan, Sung, Cheng, & Chang, 2015). Results showed that mind-mapping and drawing before writing increased grammatical knowledge significantly, and were particularly appealing to younger learners. Learners reacted positively to their use overall, and displayed awareness of why they had been engaging in such activities.

Included in Fitzpatrick et al.'s (2018) review, Tsiriotakis, Vassilaki, Spantidakis and Stavrou (2017) report a fascinating study in reducing learner anxiety about the writing process through creating a structured suitable environment which rendered writing instruction more effective. Although the treatment and control groups of grade 5 and grade 6 Greek EFL learners did not usually perform specific writing tasks as part of their language learning activities, both undertook a 15 week treatment in which they carried out writing of different genres, story and expository. In one group, learners received cognitive and metacognitive strategy instruction (Self-Regulated Strategy Development, SRSD) along with other information and advice on writing skills. In the other group, they received no strategy instruction and followed the normal curriculum. The findings showed that anxiety about writing fell significantly for the intervention group, whereas it remained constant for the control group. The authors consider that developing student critical awareness and ability to learn in a social environment, for example, will better equip them to write independently later on.

The research described above on approaches to teaching writing can be summarized as follows.

Combining literacy and oracy teaching can be beneficial, especially in mixed attainment classes.
Technology has a potentially important role to play in writing development if carefully implemented but it cannot be assumed to be advantageous merely by its presence; its use must be informed and planned.
Learners require training in using IT-based information for writing.
Use of IT requires specific classroom management techniques.
Pupils with limited technological skills must be appropriately supported. Interactive writing tasks promote proficiency and engagement.
Individual differences between learners indicate the need for a wide variety of activities and strategies to develop writing skills.
Strategy training (e.g. planning, morphological awareness, cognitive and metacognitive strategies) can improve writing performance.

3.2.4 Effective approaches to teaching speaking and listening

3.2.4.1 Speaking skills

A number of studies on effective approaches to developing speaking skills are reported by Fitzpatrick et al. (2018) under the umbrella term ‘Interaction and negotiating meaning’. These include studies involving CALL, task-based learning, CLT, and games-based learning.

In a secondary school context, in a small (n=28), one-week study of Turkish learners in an American high school, Arslanyilmaz (2013) found that task-based CALL (with meaning-focused objectives) was more effective than form-focused CALL (with objectives focused on word form) for improving production in terms of fluency and accuracy, though not complexity. In a four-month long study reported earlier, Ho and Binh (2014), found that communicative grammar teaching improved secondary students’ oral production more than a Grammar Translation Method (GTM). The students in the CLT group, perhaps predictably, performed significantly better in an oral post-test than the students taught using a GTM.

Working with younger learners (aged 7-11) from the perspective of Task-Based Learning, and specifically on how type of task repetition affects speaking, García Mayo and Imaz Agirre (2016), found that there was no difference in the effect of the type of task in terms of whether it was task repetition (with the same task and content) or procedural repetition (with the same task and different content) on negotiation of meaning strategies (such as use of clarification and confirmation checks). However, they did find that procedural repetition had the effect of creating more collaborative interactional patterns amongst learners.

Two studies at primary school level are reported which highlight the opportunities for meaningful interaction through carefully targeted games which help develop speaking skills. In their study of Greek primary school learners (mean age 7.41 years), Griva and Semoglou (2012) found that participating in physical activities and role-play games provided a real reason for children to use the target language. The children taught through this approach were more productive in communicative activities including both listening and speaking than the experimental group where a PPP (presentation, practice and production) teaching approach was used. Also working with primary school learners, Young and Wang (2014) compared the different effects of drill-based or game-based CALL pronunciation practice, and found that learners’ pronunciation showed a significant improvement when using the game-based platform. One key advantage of the game-based activities over the drill was that they provided opportunities for lively interaction with peers.

Regarding the effectiveness of computer-mediated interaction for developing speaking, Satar and Ozdener (2008), working in the context of a Turkish secondary school, showed that synchronous computer-mediated communication, whether voice chat or text chat, had positive effects on both spoken production and anxiety levels. They considered that text chat may suit lower level learners due to more time being afforded to construct utterances while higher proficiency learners can gain more from voice chat to improve fluency. For younger learners, robots with speech recognition software can be an effective means for students to improve conversation skills in a foreign language and can create a lively and positive environment where learners are free to interact and make mistakes. Wu, Wang and Chen (2015) investigated the effect of meaningful communication with a single teaching assistant robot in a class, which used forms of total physical response, communicative language teaching and storytelling techniques. They reported improvement in learning outcomes (as well as motivation/confidence) of children (aged 8-9) who worked with the robot compared to a control group. In another study, Wang, Young and Jang (2013) report the benefits of primary-school aged children (7-8 years old) working with robots ('tangible learning companions') that have the appearance of soft toys and can respond in simple conversations in English.

In updating our review of research on speaking skills, we have added a study (rated 3* overall) by Van de Gughte, Rijlaarsdam, Braaksma and Bimmel (2019) which investigated whether the type of pre-task planning can make a difference to guided peer-video modelling spoken task performance. They randomly assigned 48 ninth grade Dutch learners of German at A2 (CEFR) level (mean age 14.2 years) to two different planning conditions: short video clips of two girls showing prospective students around their school cafeteria with either a pre-task planning focus on language (FonL) (prepositions + dative case) or with a pre-task planning focus on content (FonC) (appeal and persuasiveness of presentations). The girls in the videos performed the task accurately and with confidence and before watching the two videos, the participants were informed that they were going to perform a similar task as in the videos afterwards. They watched the video clips individually in the school computer lab having been given pre-task planning sheets according to the condition. They had control over viewing and could pause and rewind the clips. Both groups watched the clips for an average of about 22 minutes. In the pre-test, two weeks before the intervention, participants were asked to describe the school cafeteria orally to prospective students to inform and persuade them to enrol in their school. Each participant then performed the same task as in the pre-test in a separate room after viewing the video clips. They performed a similar version of the main task three weeks later. Students in the FonL condition outperformed those in the FonC condition on both attempted and accurate use of the target structure on the initial post-test but this difference disappeared in the delayed post-test. Conversely, students in the FonC condition outperformed those in the FonL condition on the amount of coordination in the initial post-test and on the subordination measure in the delayed post-test. While there were some trade-offs for accuracy and complexity according to the focus of the pre-task, the authors conclude that, depending on the purposes of the lesson, the observation of peer-model videos with different planning foci can be motivating and can be effectively used to promote the accurate use of targeted grammatical structures and improve complexity during subsequent task performance.

Among our 2*-rated studies we find some further evidence that approaches that enable meaningful interaction are important to the development of speaking skills. For example, Owen, Razali, Abd Samad and Noordin (2019) investigated the effects of particular Communicative Language Teaching (CLT) activities (information gaps and language games) on L2 English speaking performance among Libyan first-year secondary school students. In-tact classes were randomly assigned to one of four groups: (1) a Language Game group; (2) a Language Game and Information Gap group; (3) an Information Gap group; and (4) a Control group. Analyses of scores on a pre- and post-test of ‘oral speaking’ found no difference between groups at pre-test, and that only the scores of Groups 3 (Information Gap) and 4 (Control) were significantly different at post-test with a medium effect size. Likewise, Winasih, Cahyono and Prayogo (2019) explored the effect on L2 speaking of combining Problem Based Learning (PBL) with e-poster creation against a ‘traditional EFL approach’ among secondary school pupils in Indonesia. Treatment was administered over a period of four classes, in which the experimental group completed problem-solving group work using e-posters, and the controls created a text in groups. An L2 speaking measure was administered pre and post the treatment phase, scored on content, vocabulary, pronunciation, accuracy, and fluency. Results indicated that there was a significant difference between groups at post-test, with the experimental group outperforming the controls. The authors explain this by saying "the speaking activities offered through learning stages of PBL were proven effective to accommodate students in increasing their speaking ability" (p. 82); however, the experimental condition included more speaking activities than the control condition, which may instead/additionally explain the difference found.

The dearth of studies on speaking and listening carried out in a UK context is striking. Even more striking, perhaps, is that, while there are hundreds of language labs and thousands of computer labs installed in UK schools, we are aware of only one largely qualitative study of language lab use in UK secondary schools (Stockman, 2015). Given the quality of multimedia labs today, her findings indicate a failure to exploit the potential of technology-enhanced language learning: "Interestingly, productive skills take the backseat in the language lab. There is less speaking practice than one would expect, in comparison to listening activities. This is in direct conflict with the advances of technology, which make multimedia recording far easier, and of greater quality." (2015, 214).

3.2.4.2 Pronunciation

While there are no studies reported which explicitly involve the use of language labs, there are several included by Fitzpatrick et al. (2018) which report effective approaches to training learners in pronunciation using computer labs. An important aspect of all the CALL-based studies is the opportunity for learners to control and repeat tasks.

For example, Neri, Mich, Gerosa and Giuliani (2008) report a study in which 11-year-olds were trained in the pronunciation of individual words using a computer-based automatic speech recognition (ASR) system and achieved, in half the time, the same results as those who were trained by a teacher. They also mention the benefit of learners giving their full attention when using the ASR programme. Young and Wang (2014) also investigated the effectiveness of ASR software comparing a dual game- and drill-based system which enabled 7-9 year olds to repeat a level if word-level pronunciation was incorrect. The game-based activities were shown to improve the children's pronunciation over the drill activities, while the dual system was important in allowing students of different levels choice in their study and the ability to return to material they were unsure of.

Primary aged children using CALL for a read-aloud task showed Improved precision in pronunciation in a study by Nutta, Feyten, Norwood, Meros, Yoshii and Ducher (2002). The children worked with an interactive story and used the computer to record their own versions of the narration. They were able to use the recording function on the computer to pause and re-record their output, leading to more repeats and greater precision. The interactivity and provision of immediate feedback available with computer-enhanced instruction changed students' method of approaching the text and the language tasks which included listening, speaking, reading, and writing.

In a non-CALL environment focusing on training in lexical stress with Korean High School English language learners, Jung, Kim and Murphy (2017) compared collaborative priming tasks: task repetition (with the same procedure and content) and procedural repetition (repetition of the same task but with different content) and found task repetition was as effective in an immediate post-test on the production of lexical stress as procedural repetition. However, in the delayed post-test, the task repetition was twice as effective as the procedural repetition. The authors suggest that hearing and producing the same words containing the target-stress patterns repeatedly, coupled with collaborative meaning-focused tasks, aided the learners in improving lexical stress, and that knowledge of word stress might be more fully proceduralized and automatized through repeated practice with the same words.

3.2.4.3 Listening skills

Listening is often regarded as the 'Cinderella' of the four skills and the relative paucity of studies on effective approaches to teaching listening skills confirms this view. Fitzpatrick et al. (2018) report several studies which provide some evidence of how supported authentic material, adapted materials and moderated teacher input can all aid comprehension.

As reported above under effective approaches to vocabulary teaching, there is a wealth of evidence with older learners (adults and students in HE) to support the use of well-selected and graded captioned video material (i.e., with same-language subtitles intended for the deaf and hard-of-hearing) for developing listening skills. Vulchanova, Aurstad, Kvitnes and Eshuis' (2015) study of secondary school students watching English language animations with captions demonstrated that captions in the target language enhance comprehension skills more than L1 subtitles.

In our updating of Fitzpatrick et al. (2018), we found two studies which investigate the contribution of captioned viewing to listening comprehension. Teng (2019c) conducted a study (n=182) in Hong Kong with grade 6 pupils (mean age 11.57) roughly equally divided into three groups comparing the effects of full captions, 'keyword' captions (displaying those words deemed important for understanding the meaning of a sentence) and no captions on global and detailed comprehension acquisition watching two specially-prepared 10-minute video clips from a series of English video stories for young learners. The fully captioned videos were more effective for high proficiency learners' video comprehension, including global and detailed comprehension but there was no significant difference between fully captioned videos and keyword captioning videos for learners with low proficiency. Both captioned forms produced better results than the videos without captions. Both global and detailed comprehension were much greater when videos were watched with captions twice and when words were shown more frequently in the keyword condition. Chen, Liu and Todd (2018) in a study focusing on reading with the aid of captioned material (reported above under Vocabulary) suggest that reading with captions can assist with 'phonological recoding', thereby enabling learners to tune in to spoken L2 material.

At the level of listening as part of interaction and the value of pre-listening activities, Fitzpatrick et al. (2018) include four studies. Cabrera and Martínez (2001) investigated strategies used by teachers to modify spoken input in the primary classroom and found that while input from the linguistic point of view was important, improvement in comprehension required simultaneous interactive adjustment, such as repetitions, comprehension checks and gestures. This theme of interactivity as well as modification in terms of linguistic input for young learners is further supported by Verdugo and Belmonte (2007) who found that though the use of multimedia stories did improve learners' (aged 6) comprehension, the stories need to be of a slow enough pace for learners.

Studies included on pre-listening activities indicate some types of pre-listening activity are more effective than others. For example, brainstorming is often used as a pre-listening technique and Li, Wu and Lin (2017) found that 14-16 year old students who used a collaborative brainstorming technique for prediction outperformed the control group who did not. In terms of the type of brainstorming activity, those students who used picture brainstorming in the first activity did better than those who used brainstorming with words. With a similar age group, Rouhi, Nabavi and Mohebbi (2014) also found that topic preparation was effective in aiding comprehension, as was repeated listening, though previewing questions did not seem to have any effect on listening comprehension scores.

3.2.4.4 Speaking and listening strategies

There have been many studies of listening strategies over several decades which have involved HE students or adults but few which focus on school-aged children. A study not included by Fitzpatrick et al. (2018) but highly relevant to the present REA, in spite of falling outside the age range criterion, is Graham and Macaro's (2008), in which they measured the effects of strategy and skill instruction on both the listening performance and self-efficacy of 68 16-17 year-old lower-intermediate learners of French in England, against a comparison group, also comparing the effects of high- and low-scaffolded interventions Results suggested that the programme improved listening proficiency and learners' confidence about listening. Fitzpatrick et al. (2018) include Jerotijević-Tišma's (2016) study, which reports beneficial effects of metacognitive strategy (e.g. planning and evaluation, problem solving, directed attention, mental translation) instruction (SI) on listening comprehension and on the metacognitive awareness of intermediate level 14-16 year olds after five weeks of training. Also working with secondary age children, Harris (2007), previously reported above, in preliminary findings of a study on strategy interaction found SI had a positive influence on performance and motivation. However, the study suggests that SI needs to be tailored effectively to ability level, as students found the listening strategies they were taught to be complex. Also reported above, Nutta et al. (2002), working with primary age (4-6) children learning Spanish in the United States, found that while the CALL approach they were investigating showed no significant learning gains for students in terms of language production, what did improve, as a result of the opportunities the CALL approach afforded the children, were language learning strategies. Learners spent longer perfecting their answers when reading aloud, highlighting that as well as teaching language strategies, providing opportunities for learners to develop language learning strategies can have beneficial effects on output.

The evidence on effective approaches to teaching speaking and listening skills can be summarized as follows.

Approaches that create meaningful interaction more positively affect speaking skills than activities where the focus is on form.
Game-based activities (in contrast to e.g. drill-based) have a positive effect on pronunciation and communicative achievement.
Online interaction and interaction with robots can be equally as effective as face-to-face interaction in creating an effective learning environment for spoken interaction.
Opportunities to interact and to make language errors are important to the development of speaking skills.
Use of automatic speech recognition (ASR) technology to provide feedback aids pronunciation and is more efficient but not necessarily more effective than a teacher.
Task repetition aids pronunciation.
Supporting authentic listening material (even with L2 input in a different medium) and moderating input both aid comprehension.
Viewing appropriately selected and graded L2 video material with or without L2 captions, either with teacher guidance or under learner control, can help to improve listening comprehension, listening skills and phonological recoding.
Pre- listening support aids comprehension.
Strategy Instruction or providing opportunities for students to develop strategies for speaking and listening improves both competences.

3.2.5 The effectiveness of general methods and approaches to language teaching

Fitzpatrick et al. (2018)) include 20 items in their evidence synthesis on effective methods and approaches for developing general language competence in the L2, as well as a number of studies which were in their 'grey literature' in the form of commissioned pedagogic reviews and reports. In our updating we found no studies which met our criteria under the theme of 'General language competence'.

According to Fitzpatrick et al. (2018), the main conclusions of these reviews (e.g. Bauckham's (2016) modern foreign languages pedagogy review concentrating on teaching practices in

Years 7, 8 and 9 in maintained schools in England and Wales) and Edelenbos, Johnstone & Kubanek's (2007) report on early language learning) do not point to any particular methods or approaches or the most effective way of teaching; so much depends on the skill and competence of the teacher delivering the teaching irrespective of method. Our view is that this position certainly reflects the reality of language teaching and learning not only in the UK but all over the world.

Given that textbooks still provide a major contribution to effective language teaching and learning, it is surprising that neither Fitzpatrick et al's (2018) review nor our own updated review found any substantial research on the quality, content or value of textbooks; yet, as Bauckham (2016, 18) reported, "Well-constructed textbooks can be a very important support both for teachers and pupils. For teachers, they embody a pattern of progression that forms a centrepiece for a scheme of work, and for pupils they can create a reference and revision guide, particularly where they are allowed to take them home for practice and revision…….. However, not all textbooks in current use are sufficiently well constructed or comprehensive to form the basis of a whole course."

Bauckham's critical comments on the weaknesses of some current MFL textbooks contrast with his positive reference to a series of Latin textbooks with a strong narrative about life and culture in Roman times (2016, 13). Indeed, in an article in the *TES* in 2012, Vanderplank contrasts the strong narrative and personal story of a Roman boy which runs through the Cambridge Latin Course with the jumble of 200 people (or rather names) in the typical Key Stage 3 textbook he describes (Vanderplank, 2012).

3.2.5.1 Amount and distribution of instruction time

Is the intensity and timing of foreign language teaching a critical factor in effective learning as well as the total amount of class time? Fitzpatrick et al. include four studies on this theme and we have also included the final version of Michell and Myles (2019) report, which was only available in 'forthcoming' form for Fitzpatrick et al's review. The findings of studies which compared intensive language training with more distributed training are mixed and inconclusive but offer insights into the handling of time distribution in the context of MFL teaching in England. In Harris and Ó Duíbhir's (2011) report (reported at length below), the main finding is that shorter-term intensive language programmes are more effective than ‘drip feed programmes’ taking place over a longer time period, as is also the case in Collins, Halter, Lightbown and Spada’s (1999) study of English language learning gains in a Québec primary school in two time distribution conditions: 2 hours of language class per day for 10 months, and the same number of contact hours condensed into 5 months (n=700). Language gains were evident in all learners, but those on the intensive programme (which was much more like an immersion programme as a variety of subjects were taught in English) scored higher across all language skills tested. However, in a better controlled and far more nuanced replication study by Collins and White (2011), they examined the acquisition of English by 11-year-old French L1 learners in Québec (n=230) in two 400-hour programmes, one delivered across 10 months and one concentrated into 5 months, comparing language development across the two contexts four times via a battery of comprehension and production measures, Both groups improved, with no significant difference in learning outcomes between the groups. Specific aspects of instructional practice, such as homework viewing of English language TV programmes by the intensive group and focus on preparation for end of year tests by the distributed group, probably led to differential effects on certain skills such as listening and on the results of some tests. For example, the TV homework feature of the concentrated students’ intensive experience resulted in extra practice with listening comprehension which may have contributed to the better performance of these students on the listening tasks at Times 3 and 4.

Fitzpatrick et al. (2018) include a ‘forthcoming’ article by Mitchell and Myles (2018) which was subsequently published (2019). Their report on 38 observed hours of language classes (no comparison group) across one school year, for 7-8 year-old English L1 learners of French (Year 3, n=26) makes sobering reading. The classes used role-play, stories, songs and crafts, following a largely oral approach. Post-tests were a receptive vocabulary test (based on classroom input) and an Elicited Imitation (EI) test to measure general proficiency. Significant gains were made in the Elicited Imitation Test, but not in vocabulary. Not surprisingly, the authors conclude that a time allowance of 38 hours per school year makes appropriate language gains challenging, regardless of the teaching approach and quality. Perhaps more important for the present REA is their conclusion that behavioural and emotional engagement are not by themselves sufficient to promote successful classroom learning and that, if they are to succeed, it is necessary even for young children to take some responsibility for managing their own learning from moment to moment. While it may be tempting to maximise the proportion of available time spent in direct learning activities, Mitchell and Myles (2019) consider children need to familiarised with target language sounds, building vocabulary and formulaic competence as well as taking initial steps in developing productive morphosyntactic knowledge.

3.2.5.2 Isolated and integrated Form Focused Instruction

Is form focused instruction (FFI) (defined by Ellis as "any planned or incidental instructional activity that is intended to induce language learners to pay attention to linguistic form" (2001, 2)) an effective teaching approach? Fitzpatrick et al. (2018) report several reviews which compare integrated or isolated approaches to FFI, that is, teaching grammatical and lexical features in isolation or integrating within a communicative, content-based or task-based approach.

Following a review of the evidence on the comparative effectiveness of these approaches, Spada and Lightbown (2008) concluded that each had a different role to play depending on factors such as the L1 of learners, rule complexity, salience of the linguistic form, its communicative value, teacher and learner preferences, and the age and development stage of learners. Isolated FFI lessons might usefully focus on language elements which were known to be challenging (possibly due to L1 influence), but were likely to be most beneficial when included in a programme of study alongside communicative and content-based classes. Integrated FFI, which might range from responsive feedback to planned, repeated elicitation of target language structures, can boost reliable usage of recently acquired language.

Without specifying age thresholds (since there are likely to be developmental factors) Spada and Lightbown (2008) also report research evidence to suggest that older children will derive more benefit from FFI, whereas young learners sometimes need little or no FFI instruction. However, in a study comparing isolated and integrated FFI with 11-12 year old learners in a primary school in Turkey, Elgün-Gündüz, Akcan and Bayyurt (2012) found that learners receiving integrated FFI performed better in vocabulary, grammar, and writing development measures than students receiving isolated FFI.

3.2.5.3 Technology and language learning

What is the evidence in support of the use of technology in general for effective language teaching? There are many research studies included in our assessment which harness various forms of technology in assisting the development of specific skills and these were reviewed above. In this section, we include two meta-analyses from Fitzpatrick et al.'s (2018) review which have broader implications for effective teaching and learning.

Boulton and Cobb (2017) provide a systematic meta-analysis of 64 studies on the effectiveness of a corpus linguistics approach to second language learning, either compared to 'traditional' approaches as a control condition or as a 'stand-alone'. They found that the greatest gains happen when learners operate a concordance program themselves, directly or through a CALL (Computer Assisted Language Learning) programme. Data Driven Learning (DDL), described by the authors as using the tools and techniques of corpus linguistics for second language learning or use, was found to be beneficial for learner language development, with large effect sizes reported for both between- and within-group comparisons (d = 0.95 and d = 1.50, respectively). The authors conclude that while the approach may be time-consuming for learners, it does appear to help in the development of language sensitivity, noticing, and inductive skills, and encourages autonomous learning and engagement with authentic language.

Grgurović, Chapelle and Shelley (2013) focus on more traditional CALL methods in their meta-analysis of effectiveness studies on computer technology-supported language learning from 1973 to 2006. They include 37 studies comparing approaches supported by and not supported by computer technology. Second/foreign language instruction supported by computer technology was found to be at least as effective as instruction without technology, and in studies using rigorous research designs the CALL groups outperformed the non-CALL groups. The results of this study highlight the importance of well thought-out and appropriate use of technology-supported language learning with experienced teachers and well-trained learners – a valuable tool when well used, a waste of time and valuable resources when not.

We summarise the main findings from Fitzpatrick et al (2018) on general language competence as follows.

Short-term intensive language programmes may be more effective than programmes where time is distributed over a longer period, but there are likely to be trade-offs and the evidence is inconclusive

Form focused instruction should be strategically deployed, and may be more beneficial to older children.

Data driven learning such as training learners to use corpora for contextual word searches can develop language sensitivity, noticing, and inductive skills, and encourages autonomous learning and engagement with authentic language.

In general, considering the evidence from all six sections, approaches and methods can be less influential on learning than factors such as teachers' language confidence, opportunities for language exposure, societal and educational context for learning, effective teacher training and ongoing skills development, motivating learners and making the language relevant to them in terms of their interests, community and identity.

3.2.6 Populations, settings and study designs

As will have been noted in the above summaries, many of the studies reported involve learners of English as a foreign language in school settings where the expectation is that educated, professional life will frequently require English language knowledge and skills. The studies included in this review which involve EFL are from China (including Hong Kong), Japan, South Korea and Taiwan, Vietnam, Spain, Israel, Malaysia, Romania, The Netherlands, Italy, Norway, Iran and Turkey. Included research into the effectiveness of approaches for teaching other languages is far more limited and is set in the UK (French, German), The Netherlands (German, Spanish), Australia (Italian), Romania (Italian). Most studies, both L2 English and other L2 languages, are set within a primary or secondary school context and cover all school ages up to 17. A number of studies were carried out in computer labs but there are no reported studies involving language laboratories. There are no studies comparing textbook use. The most favoured design involves one or more treatment groups and a control group, with pre-tests, post-tests, and often a delayed post-test. While statistics to establish equivalence of groups at pre-test and significant differences in outcome scores on post-tests are reported, few studies report effect sizes and researcher or other forms of bias are often apparent though not reported. Many studies report positively on technological support, predominantly the use of computer labs, for training in grammar, pronunciation, oral production, reading and writing, *Putting the evidence in context for England*

The context of primary and secondary MFL teaching in England inevitably limits the relevance of much of the evidence reported in this REA. Many of the studies, especially those involving EFL are set in contexts where there is not only both easy and expected access to English language sources but also an understanding that English knowledge and skills will be a valuable asset to learners in their future lives and careers. However, the evidence presented has much to offer in the way of insights, innovation and examples of effective practice which could be implemented in the English context.

3.2.7 Conclusions to Fitzpatrick et al. (2018) and updated studies

The evidence from the Fitzpatrick et al. (2018) REA together with the evidence from our updated studies provides a rich global array of insights, innovations and examples of effective approaches and practice, though we were disappointed not to find more high quality and trustworthy studies of effective approaches and good practice within the MFL context in the UK. While there is relatively little that is ground-breaking, the findings do provide a strong evidence base for changes in practice in the future. For example, our evidence indicates that approaches and methods which incorporate speaking, reading and writing at primary level can both be effective in developing grammatical knowledge and awareness as well as writing skills. At the same time, individual developmental differences in ability to handle linguistic concepts should be allowed for in primary school level teaching. Form-focused Instruction and Processing Instruction as methods for teaching grammar can be effective for teaching grammar, especially difficult concepts, but are likely to achieve better outcomes with KS3 and KS4 learners than younger learners. Both approaches are likely to be effective if there is contextualized input, the occurrence of negotiation of meaning, and student-initiated production.

Our updated studies also revealed a number of instances of effective practice in using a variety of technologies for language learning and teaching. When technology and associated material are well-targeted for specific activities and both teachers and learners are well trained in the use of various forms of technology, the evidence is clear that learning can be just successful as with teacher-led activities; where learners have control of the technology , it can be even more successful and encourage learner autonomy, language awareness, self-regulation and engagement with authentic language. For example, training learners to use the record and playback functions in computer labs is an effective method of improving accuracy in pronunciation. The research-led use of well-selected and graded L2 video clips and TV programmes, especially with L2 captions and some guidance from teachers, has long been an important feature of adult foreign language learning, especially in EF/ESL contexts. Now, our updated studies provide evidence at both primary and secondary level of their effectiveness in learning new vocabulary and developing listening and reading skills, as well as aural word recognition and "phonological recoding" of foreign sounds, particularly where learners are able to control the pace of viewing in computer/language labs.

The evidence on effective approaches to teaching reading skills presented by Fitzpatrick et al. (2018) and our updated studies confirms the value of explicit attention to both lower-level and higher-level processes, along with instruction in reading strategies and skills. Teachers should also take learners' proficiency level into account when adopting extensive or intensive reading approaches to vocabulary development. Phonics instruction (decoding) can also help beginner learners to process word forms and improve pronunciation. The evidence also indicates that technology-supported reading, including computer-assisted learning of reading, can effectively contribute to the development of oral reading fluency by facilitating student-centred learning, reducing teacher dependency and assisting lower ability learners. For writing skills, we found that effective approaches to teaching L2 writing skills require a wide variety of activities, including interactive writing tasks, and explicit teaching of strategies (e.g., planning, morphological awareness, cognitive and metacognitive strategies) in order to take Individual differences between learners into account.

There are several areas where the evidence confirms established approaches and effective practices. For example, providing opportunities for students to develop strategies for speaking and listening is an effective method for improving both skills. Methods using engaging tasks and game-based activities that create meaningful interaction are more effective for the development of L2 speaking skills, pronunciation and communicative achievement than activities where the focus is on form. While methods which encourage meaningful interactive activities are, unsurprisingly, effective for developing speaking skills, the evidence also indicates that they may be just as effective when conducted on-line as face-to-face.

In general, however, drawing on Fitzpatrick et al.'s (2018) findings and those of our updated studies, the evidence-based response to the question of effective teaching and learning of MFL in England is, perhaps, more about 'who' than 'what'. Our rapid evidence assessment indicates that 'effectiveness' lies in the hands of the professionally competent and linguistically knowledgeable teacher who is experienced in a variety of approaches, methods and strategies, is able to make flexible and pedagogically-sound use of high quality materials which are level and age-appropriate, and can draw on a wide range of technologies to best effect through training learners well in self-regulating use.

The key findings and implications of our update to the Fitzpatrick et al (2018) seed review are combined with those from our update of Harris and Ó Duibhir (2011) seed review and are presented in tabular form in Table 3.3, after we have discussed the Harris and Ó Duibhir review update. This is because we see RQ3 (addressed by the Harris & Ó Duibhir update) as subsidiary to RQ1 (address by the Fitzpatrick et al update).

3.3 Update to the Harris and O'Duibhir (2011) seed review

Background

Harris and Ó Duibhir's (2011) synthesis of research evidence was commissioned by the Republic of Ireland's National Council for Curriculum and Assessment (NCCA). The aim was to "identify, evaluate, analyse and synthesise evidence from Irish and international research about language teaching and learning in order to inform discussion about language in the Primary School Curriculum, and in particular the teaching of Irish and additional languages" (p. 12). In other words, the review set out to identify key principles for successful classroom-based language teaching, "in contexts similar to primary schools in Ireland" (p. 13). Given the context of informing discussions about the primary school curriculum, it is unsurprising that the review focuses not only on teaching practices that a classroom teacher may consider adopting, but also larger-scale policy decisions to be taken at the level of the whole school or education system.

The phrase 'in contexts similar to primary schools in Ireland' raises interesting questions about the balance between universals of SLA and context-specific issues, and the extent to which evidence of best practice can be transferred across different teaching contexts (a question that will be returned to below under the theme of teacher professionalism). A possible concern which arises when reading Harris and Ó Duibhir's review is that, in searching for evidence which is as directly relevant to the Irish primary school context as possible, much other potentially relevant evidence has been excluded.

What are the distinctive features of language teaching in Irish primary schools? As the authors point out, Ireland has "a complicated linguistic landscape" (p. 19), which is also distinctive in some respects: for example, the fact that Irish is a minority language whilst also being one of the country's two official languages. The authors also note some other key features of Irish teaching in primary schools in Ireland: first, Ireland has a long-established tradition of teaching primary school Irish lessons largely through the medium of Irish; second, the L2 (Irish) is usually taught by class teachers rather than visiting specialists; and third, primary teachers must demonstrate satisfactory competence in Irish as a condition of gaining qualified teacher status; and fourth, it is not uncommon for some other subjects to be taught through the medium of Irish, substantially increasing pupils' engagement with the target language. Some of Harris and Ó Duibhir's (2011) recommendations – for example, those relating to CLIL and immersion teaching – need to be viewed in light of these particular contextual factors. The context of primary school foreign language teaching in Ireland is rather different from (and more favourable than) that experienced in some other jurisdictions, including the UK.

It should also be noted that Irish is taught in at least two distinct contexts in Ireland, both of which are addressed in Harris and Ó Duibhir's review:

1. So-called "core" second language (L2) programmes (or L3 in the case of migrant children) with the target language being taught as a school subject. This context relates to Review Questions 1 and 3 in the current REA, which are being addressed in this section.
2. L2 immersion settings, where Irish is used as the medium of instruction for some or all of the school day. This context relates to Review Questions 4 and 5, and so evidence from Harris and Ó Duibhir's research synthesis (and from its update studies) which are relevant to these questions is covered elsewhere in the current REA (see section 3.3).

Methods

The authors conceptualize their report as a form of 'Best Evidence Synthesis, though without the involvement of an expert advisory group. The review had two phases. The first (main) phase involved a systematic search for research studies, relevant to the teaching of Irish L2 in Ireland, which (a) clearly and precisely described the instructional practices under investigation, and (b) provided evidence for the effectiveness of those practices. The inclusion criteria, quoted directly from the original report (pp. 29-30) were as follows. Included sources had to:

1. involve learners in the primary school years (4-12 age range) or inform language teaching for these pupils
2. focus on effective language teaching and learning in a school setting within the normal school day
3. concern research studies published between 1980 and 2010
4. have a process-product type design with well-defined independent (effective instructional practices or approaches) and dependent (e.g. pupil performance, or attitudes) variables
5. relate to the language teaching and/or learning in one of the following three contexts:
    - core second language (L2) programmes (…)
    - L2 immersion settings (…)
    - heritage/minority/regional/endangered language programmes, where the goal is language maintenance in the case of L1 pupils and language revitalisation in the case of L2 pupils.

In a second phase of the review, "the research design was modified to include an overview of principles for effective language teaching drawn from more descriptive (qualitative) data than from the quasi-experimental (quantitative data) in the original synthesis" (p. 17). This second phase was added later, because the authors felt that important evidence was being overlooked as a result of the strict inclusion and exclusion criteria. The aim was to provide a more complete picture of effective L2 teaching practice, as derived from various sources including "established professional opinion and experience", "judgements of recognised experts" (p. 60) and the wider SLA literature, including previous research syntheses (particularly Driscoll et al., 2004; Edelenbos, Johnstone & Kubanek, 2006; and Cable et al., 2010). Clearly, the methods for this phase of the review are less transparent (and, it seems, more reliant on the authors' subjective opinions), and we are unable to update the evidence by applying the authors' original search criteria (since these are not stated). Further, the seven main themes emerging from Phase 2 are discussed only very briefly, with limited evidence being presented to back up the key claims for best practice. Therefore, we do not include a detailed review of these claims in our own REA. However, several of the themes emerging from Phase 2 are covered elsewhere in this document. For reference, the seven themes are as follows:

- Early language learning
- Task-based interaction
- Balancing form-focused and meaning activities
- Listening comprehension and story-telling activities
- Target language use
- The European Language Portfolio (ELP)
- Language learning strategies

Returning to Phase 1, Harris and Ó Duibhir's systematic electronic database search (including so-called 'fugitive' or 'grey' literature, such as dissertations and other unpublished papers) identified 8359 potentially relevant studies. Of these, 532 were selected for abstract screening, which led to the selection of 76 sources for full-text review. 24 of these 76 contained empirical data, but of these, only 5 met all the inclusion criteria. Additionally, manual searching identified a further 17 potentially relevant studies. 15 of these contained empirical data, of which 8 met the criteria for full text review. Thus, all told, Harris and Ó Duibhir's review is based on evidence from 13 studies. For each of these studies, the authors assessed the 'strength of evidence' using a three-point scale (strong – moderate – weak). They acknowledge that this is a subjective process, and do not report any process of inter-rater agreement, which may be considered a weakness of the review.

Findings

Phase 1 of the Harris and Ó Duibhir review (i.e. the synthesis of evidence resulting from the systematic literature search) provided evidence for "a number of practices that can be said to be effective for second language learners in contexts similar to primary schools in Ireland" (p. 13). These were grouped into five themes:

- Corrective feedback
- Content and Language Integrated Learning (CLIL)
- Intensive language programmes.
- Orientation of language programmes (communicative versus analytical approaches) and the importance of teacher factors
- L2 literacy development

However, it must be noted that each theme is based on only a very small number of studies (n=2 or 3). Therefore, the evidence base underpinning each theme is slight. (Indeed, this is why the authors subsequently decided to embark upon Phase 2 of the review as described above, in order to provide a more complete picture of effective L2 teaching and learning). However, one wonders whether more inclusive search terms might have been appropriate in the systematic phase of the review, so as to broaden the evidence on which its findings were based.

When seeking evidence to inform practice, the question again arises of how close a match there needs to be between the 'sending' context on the one hand (i.e. the one in which a given study was conducted), and the 'receiving' context on the other (i.e. the one for which evidence is being sought; in this case, primary schools in Ireland). In the case of Harris and Ó Duibhir, it is not clear to us that the optimal balance has been struck here. For example, any studies pertaining to children older than 12 years were excluded; yet there may be studies conducted with Year 7 pupils (age 13) which have considerable relevance for teaching an L2 in Year 6 (particularly as the division between the primary and secondary phases of education is an arbitrary organizational one, rather than being based on any inherent divide between pupils up to age 12 and pupils above this age). However, we of course acknowledge that difficult decisions will always need to be made about the scale and scope of a review, based on the time and resources available to complete it. This applies equally to our own REA and so in updating the Harris and Ó Duibhir review, we retained the original search terms and inclusion / exclusion criteria, despite being aware of their limitations.

In sections below 3.1.13 to 3.1.17 below, we summarize the key findings of Harris and Ó Duibhir's (2011) review under each of the five key themes listed above. We also add relevant evidence arising through our updating of the review. Fifteen studies were identified as part of this updating process, but five of these were given a rating of zero stars when evaluated using Gorard's (2014) sieve. These five studies (Aljohani, 2016; Bavi, 2018; Nemati et al., 2017; Shi, 2018; Safataj & Amiryousefi, 2016) were thus deemed to provide minimal or no trustworthy evidence and are excluded from our REA. Two other studies (Berens et al., 2013; Figueroa Murphy, 2014) were deemed more relevant to Review Questions 4 and 5. They are therefore not included in the current section, which is concerned with Review Questions 1 and 3. This leaves eight studies, of which three fall within Harris and Ó Duibhir's theme of 'L2 literacy development' and are treated in section 3.1.17. The remaining five update studies do not fit into any of the existing five themes. (Given the small number of studies on which the original review was based, it is unsurprising that a number of studies identified in our update do not fit into the existing themes). These five studies are discussed in a separate section, 3.1.18.

Table 3.2 lists the studies that resulted from our update of the Harris and Ó Duibhir review, and which are relevant to Review Questions 1 and 3. The table also contains some information on the studies' characteristics and the trustworthiness ratings that they were assigned as part of the current REA (see section 2.3.3).

Table 3.2. Update studies to Harris and Ó Duibhir, 2011

Study	Topic	Context	Sample	Findings (including effect sizes, where given)	Trust-worthiness rating in this REA
Álvarez-Marinelli et al. (2016)	Computer-assisted language learning (CALL)	Primary schools in Costa Rica	n=816 pupils in 76 schools	Pupils following CALL Programme 'A' (involving "synchronized activation of the auditory, phonological, and visual systems in the brain") were compared with pupils following CALL programme B, with a focus on developing literacy, vocabulary knowledge, listening and speaking, and with a Comparison group who received English instruction from a teacher only (no CALL). Pupils following treatment A made significantly greater progress in vocabulary knowledge and listening skills compared to	3*

				the other two groups. Effect sizes were *g* = .32 to .40 respectively	
Balcı & Çakır (2012)	Teaching vocabulary through collocations	State primary school in Turkey	2 grade 7 classes, one with 30 pupils and one with 29, mean age 13.	The experimental class was taught new words in conjunction with their most frequent collocates (e.g. spend: time, money, energy, a weekend), whereas the control group was taught the same words using more traditional techniques (e.g. synonyms, antonyms, L1 translations). The experimental group showed significant advantages on a researcher-developed 'proficiency' test at post-test (*g*=1.15)†, and significantly greater recall of the target items.	1*
Buckingham & Alpaslan (2016)	Out-of-class speaking practice (oral homework)	Private school in Turkey	n=40, in 2 intact grade 3 classes: one with 19 pupils, the other with 21	Pupils in one class completed their homework orally during one semester, listening to audio recordings of their teacher and audio-recording their own oral responses. A comparison class completed the same homework tasks but in traditional written format. Only the oral homework group only improved significantly in their speaking scores, and significantly outperformed the control group at post-test. (*g*=0.73)†.	1*
Coyle & Roca de Larios (2014)	L2 literacy instruction	Primary school in southeast Spain	n=46 (26 boys and 20 girls), divided into 23 pairs of children, aged 10–12 from two EFL classes	Pupils completed an open-ended writing task in pairs. Half the pairs received direct error correction from the teacher; half saw model texts. Pupils receiving error correction noticed more grammatical features than those seeing model texts (*g*=1.40)†, made more grammatical changes when writing a revised version of their text (*g*=1.24)†, and produced significantly more acceptable and comprehensible output in their revised texts (*g*=1.11)†.	1*
de Zarobe & Zenotz (2015)	L2 literacy instruction	CLIL primary school in the Basque Country	2 intact classes, 25 students each, aged c. 10 years.	Pupils receiving metacognitive strategy instruction made significantly more progress in reading than the Control group. (*g*=1.21, a large effect)†. There was no significant difference between the Strategies and Control groups in the number or types of reading strategies reported.	1*
Gutiérrez Martinez & Ruiz de Zarobe (2017)	L2 literacy instruction	Primary schools in Cantabria, Spain: one 'CLIL' school and one 'EFL school.	6 intact Year 5 classes of around 25 pupils each (overall n=145)	Pupils receiving metacognitive reading strategy instruction performed significantly better than the Control group on reading post-tests, in both CLIL and EFL school settings, with an effect size of η2=.64.	1*
Shintani (2011)	Vocabulary learning through input-	Private English	n=36 Japanese children aged	Pupils receiving input-based instruction in a set of 24 new words, and pupils receiving production-based instruction in the same	1*

	based and production-based instruction	school in Japan	6–8, in six intact classes	words (i.e. involving oral language output) made significant gains in both receptive and productive knowledge of the target vocabulary, compared to a control group. The two groups performed similarly, with no significant differences between them on three of the four tests. There was a significant advantage for the input-based group on only one receptive test (at both post-test (g=0.92)† and delayed post-test (g=0.99).†	
Ziegler (2014)	Use of the European Languages Portfolio	4 German schools	n=318 pupils in grades 4-9 using the European Language Portfolio (ELP), plus n=257 non-ELP pupils in grades 5-9	In this natural experiment, pupils who used the European Language Portfolio (ELP) scored significantly higher than non-ELP pupils on self-reported measures of self-regulated learning, with effect sizes ranging from η^2=.02 to .06. Further, among the ELP pupils, those who used the portfolio more frequently (as reported by their teachers) showed significantly more positive self-regulated learning outcomes than those who used it less frequently.	1*

† Author (of this REA) calculated effect size

Corrective feedback

The authors conclude that corrective oral feedback can be effective in promoting L2 development amongst primary-age children, with prompts being more effective than recasts, which are in turn more effective than simply ignoring errors and providing no feedback. (Prompts are where the teacher indicates to the learner that an error has been made, and invites her/him to reformulate the utterance; recasts are where the teacher repeats the learner’s utterance but with the error corrected).

This conclusion is, however, based on only three studies, two of which involve the same author and dataset. Further, one study was extremely limited in scale and scope, being “a small-scale quasi-experimental investigation that examined the acquisition of one morphosyntactic form, the possessive determiners ‘his and ‘her’” (p. 38). On the other hand, Harris and Ó Duibhir note that the findings of these studies align well with those of other studies on corrective feedback in classroom contexts, which is a well-researched area (albeit only rarely in primary school settings). For example, they refer to Lyster and Saito’s (2010) meta-analysis of oral feedback studies in L2 classrooms. This obtained similar findings, even though it was based on studies conducted in rather different contexts than L2 Irish classrooms in Irish primary schools.

The authors add that, when giving corrective feedback in primary school L2 classrooms, care must be taken to preserve and nurture pupils' self-esteem and confidence. This seems a reasonable caveat, albeit one which would surely already be evident to most primary school teachers.

Content and Language Integrated Learning (CLIL)

The authors borrow Coyle, Hood and Marsh's (2010:1) definition of CLIL as "a dual-focused educational approach in which an additional language is used for the learning and teaching of both content and language', noting that in North America it is also widely known as Content-Based Instruction (CBI). A key conclusion of Harris and Ó Duibhir's review is that CLIL is an effective way of promoting pupils' L2 learning, with no detriment to either (a) the development of their L1 or (b) their outcomes in the subject area being taught. They argue that "CLIL enables learners to encounter language in context and use it for authentic communication, and challenges them to use the target language for cognitive purposes to acquire knowledge, skills and information" (pp. 14-15).

The authors note, however, that the evidence base relating to CLIL is rather slim (with few empirical studies having been conducted); there has been a tendency, they argue, for CLIL practice to race ahead, with research evidence lagging behind. Indeed, the evidence for the conclusions reached in Harris and Ó Duibhir's review comprises only four studies which were identified as being relevant to CLIL in Irish primary schools. Further, these studies have numerous limitations. Three of the four are naturally-occurring experiments with the risk of many confounding variables (most notably the fact that schools following a CLIL curriculum are a self-selecting sample which may differ in a number of ways from those which choose not to offer CLIL). Further, in one of the four studies, pupils in the CLIL group had also had more hours of English teaching prior to the start of the study, again presenting a major confound.

There is also a terminological question of whether the approaches described in these studies are really 'CLIL' (if conceived of as a programme with a dual focus on both content and language) or simply L2 immersion. Indeed, Harris and Ó Duibhir go on to note that CLIL teaching in Ireland may be seen as a form of immersion, which they say is supported by positive research evidence.

Overall, the conclusion of the review is that CLIL instruction may be an effective way to teach Irish in Irish primary schools. However, the transferability of this to other contexts (such as MFL teaching in English schools) is subject to question. As the authors note, Irish primary schools may be seen as a particularly favourable context for embedding CLIL in teaching L2 Irish, given the requirement for all primary teachers to have competence in the Irish language, and the long history of immersion education with L2 Irish being used as the medium of instruction. Further, it should be noted that implementing CLIL into the curriculum in any systematic and integrated way is more likely to be a policy decision at the whole school level (and requiring the support of parents) than something within the remit of individual classroom teachers.

We refer the reader to section 3.3 of this REA (and particularly the conclusions in section 3.3.6), which examines in more detail the effects on both content learning and linguistic outcomes of using L1 or L2 as the medium of instruction.

Intensive language programmes

Harris and Ó Duibhir conclude that there is “strong evidence to suggest that intensive programmes of instruction in a second or additional language over a short time period are more effective than drip feed programmes, where learners are exposed to limited amounts of the language over a longer period” (p. 15). This, they argue, is because such programmes help pupils to quickly achieve a basic level of communicative competence, allowing them to engage in spontaneous communication in the L2 and in turn sustaining and enhancing their motivation to learn the language. By ‘intensive language programme’, the authors mean the teaching of the L2 as a stand-alone subject (what they elsewhere call a ‘core’ L2 programme), rather than the CLIL or immersion approaches referred to above, although these could also serve a similar function in promoting rapid progress towards basic communicative competence.

The conclusions in this section of Harris and Ó Duibhir’s review are, once again, based on minimal evidence: a single study by Netten and Germain (2009). However, the authors argue that this study is robust, since it draws on the results of pupil assessments in intensive French programmes in nine jurisdictions in Canada. They further note that there is some convergent evidence from other contexts (university students) supporting the value of shorter, more intensive programmes of L2 learning as compared to longer, less intensive ones.

It is interesting to note that, over ten years ago, Macaro (2008) argued that a key solution to the crisis of motivation and uptake in English MFL classrooms (which still persists today despite various policy initiatives in the intervening time) is to ensure that pupils make “rapid and substantial progress” in the L2 in the first few years of learning the language at secondary school. In turn, he argues, “the only way that they can do this is to provide them with much greater amounts of teaching and learning in the crucial first 10 months of Year 7” – in other words, an intensive L2 learning programme like the ones being advocated here by Harris and Ó Duibhir. Once again, however, we find ourselves in the territory of school- and system-level policy decisions which would require support from the whole school community, and lie beyond the remit of individual classroom teachers.

Orientation of language programmes (communicative or grammatical/analytical approaches) and the importance of teacher factors

The question of which ‘orientation’ of an L2 programme – communicative versus grammatical / analytical – is an important one and lies much more clearly within the remit of individual classroom teachers and languages departments. It is also a matter of long-standing debate and controversy.

Drawing on two studies identified by their systematic search and meeting their inclusion criteria, Harris and Ó Duibhir find inconclusive evidence in relation to this key question. This, they argue, is consonant with the wider literature on this subject:

> The evidence from research shows contradictory results – in some studies communicative oriented courses did not result in any improvement in students’ proficiency while in others the language proficiency of learners in classrooms where experiential and communicative activities were emphasised were better than those where there was a traditional grammatical/analytical approach. The conclusion is that the link between course design and pupil proficiency is quite weak and is dependent on context. (p. 16)

They go on to argue for the importance of achieving the “right balance between communicative and analytical activities” (p. 16). However, research has so far been unable to provide clear guidance on what the optimal balance between meaning-focussed (communicative) and form-focussed (analytical) should be – and indeed perhaps it is unrealistic to expect research ever to be able to determine this balance for a particular classroom on a particular occasion, given the huge variety of contextual factors at play. The role of the teacher thus becomes crucial in achieving the optimal orientation, by carefully designing tasks in relation to the learners' needs at any given time (Mutton & Woore, 2014). In doing so, we would argue that teachers should also draw critically on the available research evidence and interrogate it in relation to their own particular context, in line with the model of teacher professionalism described by Winch, Oancea and Orchard (2013).

This view of the crucial importance of the teacher's professional competence and judgment also resonates with the findings of Graham et al. (2017) in their study of primary school learners of French in England. In comparing the outcomes of pupils taught by an oracy-based approach on the one hand versus a combined oracy-plus-literacy approach on the other, they found that teacher characteristics (such as level of experience and level of teacher education) made more of a difference than the teaching approach itself. In the primary school context, there are thus important implications here for the initial education and continuing professional development of L2 teachers.

Development of L2 literacy skills

Note that many primary school L2 curricula focus exclusively on the development of oral language skills (speaking and listening). A predominant focus on oracy has also been found in MFL teaching in UK primary schools (Cable et al., 2010). However, drawing on the findings of three studies included in their systematic synthesis of evidence, the authors advocate the inclusion of literacy teaching in primary L2 classrooms, concluding that "the development of students' L2 literacy skills supports the development of their second language proficiency in general" (p. 16). More specifically, they recommend that teachers should (a) introduce L2 literacy gradually, taking account of pupils' wider literacy development (including in L1); (b) teach L2 reading strategies explicitly; and (c) consider reading aloud to pupils, as a way of developing both their pronunciation (including stress and intonation) and their comprehension (by helping them to focus on larger units of meaning beyond the word level).

The three studies on which Harris and Ó Duibhir base these conclusions have a number of limitations: for example, Macaro and Mutton (2009) has problems of scalability, since the teaching was conducted in small groups who were taken out of their usual languages lessons; and Amer (1997) is small in scale. The third study (Drew, 2009) trials an extensive reading programme conducted both in school and extramurally, but it is questionable whether primary school pupils in some other contexts (including the UK) would have sufficient linguistic knowledge to enable them to do this.

It is interesting that these findings contrast with those of the more recent study by Graham et al. (2017), referred to above. This was a natural experiment, comparing the effects, in primary MFL lessons, of (a) an oracy-focussed programme of instruction and (b) a combined oracy-plus-literacy programme. No effect of teaching approach was found on pupils' outcomes. However, as the authors point out, the two teaching approaches may have been less distinct than intended: the 'oracy' group may have introduced some literacy activities as pupils neared transition to secondary schools, whilst in the combined oracy-plus-literacy group, "there was relatively little evidence of the higher-level literacy activities that may be needed for literacy to support learning" (p. 952). Greater effects may have been observed as a result of a more intensive and systematic programme of literacy instruction.

Overall, we would support the call for the integration of literacy instruction into primary MFL curricula. However, what is striking about the picture presented by Harris and Ó Duibhir is the limited range of evidence that is drawn upon. For example, there is a substantial body of literature on reading strategies that could have informed the discussion (including evidence from recent meta-analyses by Plonsky, 2011 and Ardasheva et al., 2017, as well as the systematic review by Hassan et al., 2005). Whilst these do not specifically target primary school children, we would suggest that they do contain evidence which is relevant for informing practice in primary languages classrooms. Further, there are omissions in the topics covered within the umbrella of 'literacy'. For example, anecdotal evidence suggests that the teaching of phonics is gaining popularity in UK MFL classrooms, at both primary and secondary school levels, with the teaching of phonics being strongly advocated by organizations with governmental backing (e.g. Bauckham, 2016; the National Centre for Excellence for Language Pedagogy, https://ncelp.org). To our knowledge, evidence for the effects of phonics teaching in an L2 (whether at primary level or beyond) remains relatively scarce. However, a recent cluster RCT by Woore et al. (2018) provides support for the teaching of L2 phonics to beginner learners of French, as part of an integrated package of L2 reading instruction. Whilst this study was conducted with Year 7s, it is likely that the conclusions apply equally well to primary school MFL learners in Year 6 or below.

3.3.1 Update studies on L2 literacy instruction

Our updated search (using the same criteria as Harris and Ó Duibhir's original review) revealed three additional studies which pertain to the category of literacy instruction at primary school level. These three studies (Gutiérrez Martinez & Ruiz de Zarobe, 2017; Zarobe & Zenotz, 2015; Coyle & Roca de Larios, 2014) are discussed below.

The first was a quasi-experimental investigation of the effects of metacognitive reading strategy instruction in two different educational settings: CLIL and EFL (Gutiérrez Martinez & Ruiz de Zarobe, 2017). (The comparison of the two contexts means that this study is also relevant to Section 3.3 of this REA). Two different schools in Cantabria, Spain, took part. One was a CLIL school, whose Year 5 pupils had three hours of Science and one hour of arts/crafts taught in English each week, in addition to three hours per week of EFL lessons. The second was an 'EFL' school, in which pupils had much less exposure to English each week (three hours per week in discrete English lessons). The sample comprised three intact classes per school of around 25 students each (n=145 in total; age roughly 10 years; 57% ♀).

In each school, two classes were randomly assigned to the experimental group and one to the control group. Participants in the experimental group received seven sessions of metacognitive strategy instruction, delivered by one of the researchers and following the cyclical model proposed by Macaro (2001). This comprises initial awareness raising, modelling of new strategies, scaffolded practice, removal of the scaffolding, and evaluation of strategic behaviour in relation to outcomes.

Participants' outcomes were compared on pre-tests, immediate post-tests and delayed post-tests (six months after the end of the intervention). These tests are described as "metacognitive reading tests" comprising "25 open-ended questions concerning the reading strategies worked on during the training". However, on the basis of the information provided, their precise nature is unclear, particularly whether they include reading outcomes or are limited to participants' strategic behaviour.

Taking participants as a whole, the experimental groups were found to significantly outperform the control groups in both schools, with a moderate effect size (η^2=.64) based on Cohen's (1998) guidelines. No evidence was found of differential effectiveness of the training according to instructional context (CLIL versus EFL). The authors conclude that strategy use seems to be "a powerful tool in second language classrooms not only to improve learners' reading competence, but also to help them become better, more independent learners able to monitor their own learning process" (p. 86).

However, the trustworthiness of this conclusion is undermined by a series of limitations to the study. First, there is a lack of information on the nature of the teaching received by the control group, particularly with regard to strategy instruction. Second, much more information is needed in order to be able to assess the validity and reliability of the tests used; indeed, it is not even clear from the information provided what, precisely, these tests are measuring. Third, there is a risk of teacher effects influencing outcomes: the intervention groups, but not the controls, were taught by one of the researchers. Fourth, although there is random allocation of groups to conditions, the very small sample size means that this is insufficient to control for pre-existing between-group differences. The analysis also fails to account for the clustering of the individual pupil data at the class level. Due to these limitations, we assessed this study as having significant risk of bias and so it does not substantially change the conclusions reported above relating to L2 literacy instruction.

The second update study, by de Zarobe and Zenotz (2015), is similar to the previous one in that it examines the effects of reading strategy instruction in a CLIL context. Two intact classes from a school in the Basque Country took part. The classes each had 25 pupils (68% female), aged approximately 10 years. Most pupils' L1 was Spanish, but they had been learning both Basque and English since the age of four.

One class was allocated to an experimental condition and the other to a control. The means of determining which group was allocated to which condition is not stated. The experimental group received seven sessions (spanning three months in total) of instruction in reading strategies, within their usual lessons but delivered by the researchers. The strategy instruction again followed Macaro's (2001) cyclical model (see above). The intervention used texts from Science, one of the subjects that participants were being taught through the medium of English. The control group followed their usual lessons with their usual teachers, and did not receive the strategy instruction.

Pre- and post-tests were used to measure participants' outcomes. These comprised (a) a reading strategy survey, (b) a so-called 'metacognitive reading task', and (c) a metacognitive task, completed immediately after the reading task and aiming to probe participants' reading processes and strategies during that specific reading task (rather than the more general information on strategy behaviour elicited in the reading strategy survey). As in the previous study, there is some lack of clarity concerning the precise nature of the metacognitive reading task. The authors state that it "consisted of 25 open-ended questions concerned with reading skills such as 'skimming', 'scanning' and 'detailed reading'. Yet, the use of certain strategies such as 'predicting', 'guessing from the context' and 'observing the layout of the text' was also tested" (p. 324). Identical tests were used at pre- and post-test.

No significant differences were found between the experimental group (who received the strategy instruction) and the control group (no strategy instruction) in terms of the number and types of reading strategies that they reported. Neither were there any significant differences between the groups' scores on the metacognitive reading test, either at pre- or post-test. Nonetheless, the experimental group showed significantly greater progress in reading over the course of the experiment, as measured by pre-post-test gain scores. (An effect size is not reported).

The authors conclude that "the strategy intervention had a positive effect on the reading comprehension process"; thus, they claim, their study supports "the body of evidence that suggests that strategy instruction is an effective tool in second language classrooms to increase the reading competence of learners" (p. 331). However, this conclusion would have been more strongly supported, had the experimental group also shown increased use of strategies as a result of the intervention; strategic behaviour resulting from the intervention could then have been tied to improvements in reading outcomes. As things stand, the authors have to find some other explanation for the pattern of results, for example suggesting that readers in the experimental group could have developed “a better control and awareness of how strategies are used”, thus resulting in improved reading outcomes (p. 331). However, this is speculation. Further, the study has a number of other limitations which weaken its findings. For example, it appears that intact classes were allocated to the two conditions, which given the small size of the sample (two classes only) presents a serious validity threat, due to the possibility of uncontrolled pre-existing differences between the groups. The intervention group but not the control group was taught by one of the researchers, creating a risk of teacher effects. Further, no information is given on the teaching received by the control condition, particularly in relation to strategy instruction. Finally, very little information is provided concerning the validity or reliability of the instruments used in the pre- and post-tests. Therefore, overall, we assessed this study as having significant risk of bias. Again, it does not substantially change the conclusions reported in Harris and Ó Duibhir’s original review relating to L2 literacy instruction.

The third study included in our update of Harris and Ó Duibhir’s review was conducted by Coyle and Roca de Larios (2014). Their study used an experimental design to investigate whether different types of feedback (error correction by the teacher versus provision of two correct written models) resulted in (a) noticing of different linguistic features and (b) improved writing acceptability and comprehensibility. Despite the focus of the study on corrective feedback, we have included this study here rather than in section 3.1.13 above, because it relates to the written rather than oral modality.

Participants in the study were 46 children from two EFL classes in a Spanish primary school. They had been learning English for 4-5 years, with an average of 3 hours of English lessons per week following a communicative approach. The participants were placed in pairs of matching proficiency (based on routine English test results), resulting in 7 high proficiency pairs, 9 medium proficiency and 7 low.

The intervention lasted three weeks and had three stages. First (stage 1), participants worked in pairs to write a short story in response to a four-frame picture prompt, noting down any linguistic problems they encountered. Then (stage 2), half of the pairs received direct corrective feedback on these stories from their teacher (i.e. their errors were corrected), while the other half received two model texts in response to the same story prompt. They noted down any differences they observed between their original story and either the teacher's corrections or the model texts. Finally, one week later (stage 3), the children were given the same picture prompt again and asked to rewrite their story, without access to their original or the feedback.

The data analysed in the study comprised: (a) participants' original stories (written by the children working in pairs); (b) the notes they made of any problems they encountered during the composing process; (c) the notes they made when comparing their original texts to the teacher's feedback or model texts; and (d) the revised story texts, again written in pairs.

The main findings relevant to our review questions were as follows. (Note that no effect sizes are given in the report of the findings). First, at stage 2, the participants who received error correction noticed (in the sense of Schmitt, 1990) more grammatical errors in their work. Of the language features noticed by participants, 27% were grammatical for the error correction group, compared to 0% for the model texts group. This difference was found to be statistically significant. In descriptive terms, the model texts group also noticed a higher proportion of lexical features than the error correction group (83% versus 56%), but this difference was not found to be significant. Second, the large majority of participants' revisions (when re-writing their texts at stage 3) were lexical (76%). However, learners in the error correction group were significantly more likely to incorporate into their new texts grammatical features which they had noticed at stage 2; by contrast, those in the model texts group were significantly more likely to introduce new lexical items at stage 3, which they had not mentioned at stage 2. Finally, the pupils who received error correction produced significantly more acceptable and more comprehensible output in their revised texts, compared to the pairs who received the model texts.

The authors claim that their study provides evidence that "explicit knowledge, promoted through written feedback, can have an impact, at least, on improving accuracy in writing" (p. 479). Pedagogically, they conclude that both types of feedback (direct corrective feedback and the provision of model texts) can be useful in the classroom, since they seem to promote the noticing of different linguistic features: "Error correction may be more important for directing attention to grammatical features, whereas the strength of models lies in their potential to provide lexis and expressions beyond learners' current repertoires" (p. 479).

Once again, this study has a number of limitations which call into question the trustworthiness of the evidence it provides. First, the intervention involved pairs of learners, and so the effects on individuals is not measured. Second, the quality of the pupils' writing in the post-tests relates only to the revised versions of their original texts; therefore, there is no evidence concerning the longer-term effects on their linguistic knowledge, nor how this might be applied in new pieces of writing. Third, as the authors acknowledge, there is potential reactivity, in that the reporting of challenges encountered when writing (and discussing this in pairs) may have affected the process of writing, the noticing of linguistic features and the quality of the written output. Therefore, there are issues with ecological validity. Fourth, as a result of attrition, the two experimental groups were imbalanced in terms of the participants' proficiency level, with low-proficiency pairs being under-represented in the error correction group. Finally, it is not clear what kind of instruction (and, crucially, what kind of feedback on written work) participants received beyond the scope of the intervention. As the authors note, the children continued with their regular English classes during the three weeks of the intervention, and so it is impossible to know whether any differences in outcomes between the groups truly reflect the effects of the interventions. As a result of these limitations, we felt that this study carries significant risk of bias. Our view is that it does not alter any of the existing conclusions drawn by Harris and Ó Duibhir in their original review.

Nonetheless, the study is useful in highlighting the possibility of using model texts to provide feedback to L2 learners, even in the primary school classroom. In our experience, this is a little-used pedagogical approach in UK MFL classrooms (at any level). There is evidence from other contexts that it can be an effective way of promoting learners' noticing of errors in their written work (e.g. Hanaoka, 2007) – and of course from teachers' point of view, it is a much more time-efficient way of providing feedback than individualized error correction. This is significant given current concerns over teacher workload in the UK context, as highlighted by the recent governmental review of marking practices (DfE, 2016).

3.3.2 Findings of studies from the updated literature search which do not fit into the above five categories

As noted above, our updated search additionally identified five studies which do not fit within the five existing themes in Harris and Ó Duibhir's review. These are reviewed below.

Alvarez-Marinelli et al.'s (2016) study focuses on computer-assisted language learning (CALL). It is interesting in itself that Harris and Ó Duibhir's review does not include any findings relating to CALL or, more generally, to the use of technology to support language learning. The paper presents initial findings from a larger (two-year) cluster RCT investigating the effectiveness of CALL approaches for the learning of English as a Foreign Language by children in Costa Rican primary schools. The sample comprised 76 schools, randomly selected from primary schools in the country, with a total of 816 third graders participating. A large majority of schools were in rural locations, but there was also a handful of urban schools, chiefly in the comparison group.

Schools were randomly allocated to one of three conditions. In Treatment A, schools were provided with "computer-assisted English language learning software, tests, assessment tools, and training to support teachers … [A] key element is the synchronized activation of the auditory, phonological, and visual systems in the brain, especially important for listening and reading development" (p. 108). Pupils engaged with technology-based instruction for five days per week for 25 weeks, with an average time-on-task of just over 1 hour per week. In Treatment B, pupils followed a different, "research-based language acquisition CALL curriculum especially designed to meet the needs of English Language Learners" (p. 109), with a focus on developing their literacy, vocabulary knowledge and listening and speaking skills. The average time-on-task for Treatment B was just over 2 hours per week, again for 25 weeks. Finally, a no treatment comparison group followed what is described as "the typical English instruction model in Costa Rica" (p. 109), receiving English instruction from a teacher only (no CALL) for an average of 2.5 hours per week over 25 weeks. The focus in these lessons was on listening and speaking.

The effects of the two interventions were evaluated by means of pre- and post-tests comprising four subtests from the Woodcock Muñoz Language Survey-Revised, designed to measure children's oral English language development. The groups' scores were compared using a multi-level model, thus taking account of the clustering of data at the school level (because children in the same school, living in the same area, taught by the same teacher are likely to perform similarly to some extent).

The key finding of the study, for our current purposes, is that pupils who followed the CALL programme in Treatment A made greater progress in vocabulary knowledge and listening skills, compared to their peers in the Treatment B and Comparison groups. The effect sizes (Hedge's g) were .32 to .40 respectively, indicating small-to-medium effects. These effects were found even after controlling for time on task; recall that pupils following Treatment A spent the least amount of time on the intervention, half or less than the average time spent by pupils on Treatment B or the Comparison condition. Because there was an imbalance in the sample (with rural schools being under-represented in the comparison condition, and urban schools overrepresented), the analyses were repeated using the rural schools only, which comprised nearly 90% of the sample. Broadly similar findings were obtained.

The study demonstrates rigour through its large sample (whose size was determined by means of a power analysis, in order to allow a small effect size to be detected), which is able to take account of the clustered sample. The allocation of schools to conditions is also fully randomized, although this does introduce some bias into the sample. For example, as noted above, the Comparison group included more urban schools, as well as a lower percentage of pupils repeating a grade. A significant difference between the groups' performance was also found on the pre-tests, with the Comparison group scoring higher than the treatment groups. Since the analyses were based on gain scores, there is thus a possibility that the greater gains for the treatment groups were (at least partially) influenced by a regression to the mean. In another limitation of this study, whilst there is one brief mention of lesson observations "by our on-site research fidelity coordinator" (p. 111), little information is provided about what was actually done in the comparison classrooms, and how practice may have varied between teachers. Similarly, there is no attempt to document any out-of-class learning experiences that pupils may have had, including exposure to the L2 and use of technology for English learning – both of which might conceivably have differed between urban and rural schools. Thus, we deemed the study to have some risk of bias. Further, whilst there is a very general description of the CALL interventions themselves, much more detail is needed (ideally with exemplification) in order to really understand what was involved. In the absence of such information, it is difficult to judge the relevance of the study for other L2 learning contexts.

Shintani's (2011) quasi-experimental study investigated the effects of two types of treatment (one emphasizing input, the other output) on young EFL learners' vocabulary acquisition. The sample comprised six intact classes comprising 36 Japanese children aged 6-8, with two classes being allocated to each arm of the study. The first group received input-based instruction in which they covered a set of 24 target words. For example, they were given sets of cards representing target nouns. They then listened to the teacher say sentences containing those same nouns, and had to select the appropriate cards showing pictures of those nouns. The second group covered the same target words, but were required to produce them as part of their instruction. For example, they completed various tasks such as 'listen and repeat', 'guess the hidden flashcard' and 'Kim's game' (identifying the missing object from a set of objects). The third (control) group completed a set of three activities (English songs, Total Physical Response, and alphabet practice) without being exposed to the target words. Each group had two lessons per week for six weeks. Approximately 30 minutes per lesson was dedicated to each intervention, thus equalizing time on task, and all lessons were taught by the researcher in order to control for teacher effects.

At each of three time points (pre-test; immediate post-test; delayed post-test, five weeks later), participants completed four different tests designed to measure their comprehension (receptive knowledge) and recall (productive knowledge) of the target words that had been taught in the interventions. Overall, it was found that both the input-based and production-based interventions led to both receptive and productive knowledge of the target items. The two groups performed similarly, with no significant differences between them on three of the four tests. Only on one of the receptive tests was a significant advantage observed for the input-based group, on both the post-test and delayed post-test. (No effect size is given). It is interesting that there was no advantage for the production-based group on the productive tests. This may have been because the input-based tasks "provided opportunities for richer interaction for the learners than the production-based activities", as revealed through the analysis of process data (p. 137). The author concludes that "input-based tasks can be successfully implemented in EFL classrooms for young beginners and are at least as effective as production activities where vocabulary learning is concerned" (p. 156).

The author acknowledges, as limitations of the study, both the small sample size (this is particularly acute given the clustered nature of the data) and the fact that the pupils were in a private language school. The latter, it is argued, may limit generalizability to the wider population of EFL learners in Japan. Additionally, there is no indication of how groups were allocated to conditions, and the use of intact classes as the unit of allocation means that there is a risk of pre-existing differences between the groups, which may have affected results. All lessons were taught by the researcher, which reduces the risk of teacher effects, but still leaves open the possibility of bias (even if implicit), for example if the researcher were to subconsciously favour one of the interventions. Finally, no information is provided concerning the validity or reliability of the tests used. Because of these limitations, overall, we judged this study to have low strength of evidence.

Buckingham and Alpaslan's (2016) quasi-experimental study, which again falls outwith the five themes identified in Harris and Ó Duibhir's original review, was designed to investigate whether out-of-class speaking practice mediated by technology could help young learners of English improve their speaking proficiency grades. (Changes in the willingness to communicate of the experimental group only are also investigated, but are not reported here because there is no comparison with the control group on this variable). The sample comprised two intact Grade 3 classes, one with 19 children and the other with 21, randomly selected from the three grade 3 classes in a Turkish private school.

The classes were randomly assigned to either the experimental or the comparison condition. Pupils in the experimental group completed all their homework orally over the course of one semester (four months). For each homework, they received a slide presentation containing video and voice recordings of the teacher, images and text. They had to audio-record their oral responses to the teacher's questions. The control group completed the same homework tasks but in traditional written format. The teacher (one of the researchers) was the same for both the intervention and control groups.

Data from pupils' routine oral assessments (completed individually in class with the teacher-researcher, both before and after the intervention) were used to compare the effects of the two homework formats. These assessments comprised a series of 11 questions, divided into three sections: short-answer questions; description of a picture; and questions about the learner. Pupils' assessments were audio recorded and marked by the teacher-researcher and a colleague; high inter-rater reliability is reported.

There was no significant difference between the groups' oral assessment scores before the intervention. At post-test, only the experimental group was found to have improved significantly in their speaking scores, and they significantly outperformed the control group. (No effect sizes are reported). The authors conclude that "the use of computer technology to deliver out-of-class speaking practice can contribute to improving young learners' speaking proficiency scores and can positive[ly] impact on learners' willingness to communicate" (p. 34).

The study's findings are weakened by a number of limitations. First, as the authors acknowledge, the technology may have been associated with a 'novelty effect', providing a boost to pupils' outcomes in the experimental condition. Second, both groups have the same teacher (one of the researchers), which is a strength in that it controls for teacher effects; however, there remains the possibility of the teacher-researcher favouring one of the groups in either their teaching or the conduct and scoring of their assessments (which was not done blind to condition), even if only subconsciously. Third, despite the random allocation to condition, there is only one class per condition and so there is a distinct possibility of the results being influenced by pre-existing differences between the groups. Further, though not stated explicitly, it appears that the classes may have been single sex, with one being all girls and the other all boys. Finally, the authors note that the demographic of the study was socio-economically favoured relative to the broader Turkish population; thus, it may be more difficult to implement this technology-mediated intervention in other, less favoured contexts.

In view of the study's limitations, we rated it as having significant risk of bias. Nonetheless, we consider it to serve a useful function in our review, in that it highlights the possibility of different formats of homework beyond traditional written tasks. The National Centre for

Excellence for Language Pedagogy (https://ncelp.org) is advocating the use of 'spoken homework' for L2 learners at secondary school, and there is no reason why this should not also be tried with younger age groups. Spoken homework tasks may play to some learners' strengths, for example by helping to avoid some of the entrenched patterns of underachievement associated with literacy difficulties.

The study by Balcı and Çakır (2012) used a quasi-experimental design to investigate the effects of teaching L2 English vocabulary through collocations, as compared to a more traditional approach. Both of the 7th grade classes in one Turkish school took part. There were 59 pupils (30 in one class, 29 in the other) with a mean age of around 13 years. The classes were randomly assigned to either the experimental or control condition, with both being taught by the one of the researchers. The experimental class were presented with target words in conjunction with their most frequent collocates (e.g. spend: time, money, energy, a weekend). The control group was taught the same target words "through classical techniques such as synonym, antonym, definition and mother tongue equivalence" (p. 25).

To measure the effects of the teaching, a 20-item, multiple choice proficiency test (developed by the researchers, but with little further information given) was administered to all participants before and after the intervention. There were also weekly vocabulary tests based on the target words taught that week, together with a 'retention test' covering the same items after a week's delay. Whilst there were no significant differences between the groups on the proficiency pre-test, in the post-test the experimental group achieved a significantly higher score. (No effect sizes are given in this paper). The experimental group (in contrast to the control group) also showed a significant and substantial increase in their proficiency score from pre- to post-test. In the weekly vocabulary tests covering the target items that had been taught, there was no significant difference between the groups' scores for the first five weeks, but in week six, the experimental group significantly outperformed the control. The same was true of the retention test administered one week later. The authors conclude that "teaching vocabulary through collocations results in a better learning of the words than presenting them using classical techniques and enhance[s] retention of new vocabulary items" (p. 29).

As with the previous studies identified through our updated search, Balcı and Çakır's (2012) quasi-experiment has several limitations. Perhaps the most important of these is (again as in previous cases) the small sample size: with only one intact class per condition, there is a clear risk of confounding variables in the form of pre-existing differences between the groups that might affect their vocabulary learning. The fact that the classes have the same teacher controls for teacher effects, but there may nonetheless be implicit or explicit bias in favour of the intervention group. Finally, no information is given concerning the validity or reliability of the tests used to measure pupils' outcomes. Thus, we again rated this study as having significant risk of bias. Nonetheless, at a time when the importance of vocabulary teaching is being strongly reasserted (e.g. by the National Centre for Excellence for Language Pedagogy), it is useful for teachers to be made aware of alternative approaches such as the one exemplified here. The teaching of new words together with their frequent collocates can also be related to the lexico-grammar approach currently being advocated by Gianfranco Conti's 'Language Gym' website, very popular with L2 teachers in the UK and more widely (see https://gianfrancoconti.com/2018/07/30/patterns-first-how-i-teach-lexicogrammar-part-1/).

Ziegler's (2014) study adopts an 'explanatory mixed-methods sequential design' of which the natural experiment part is reported here; a second, qualitative phase (to explain the quantitative findings) is not reported. The sample included pupils from four different schools, some of which had adopted the European Language Portfolio (ELP) in their L2 English teaching, and some had not. The four schools were located in the same German city and had similar demographic characteristics, including socio-economic status. There were 318 ELP pupils in grades 4-9, with 12 different teachers, and 257 non-ELP pupils in grades 5-9, with 7 different teachers. ELP participants had begun using the portfolio in grade 3. The author does not describe what this use looks like, but did collect data on how frequently teachers reported using the ELP with their classes.

Measures of self-regulated learning were used to compare pupils in the two groups. These comprised self-reported measures of: mastery versus performance approaches to goal orientation; task value; academic self-efficacy; and strategic attributions. For the ELP group only, pupils' attitudes towards the ELP were gathered via six self-report items. These assessed pupils' beliefs about the efficacy of using the ELP, task value and enjoyment.

The findings of the study showed that, after controlling for grade level, ELP pupils reported higher mean mastery goal orientation (F=14.79, p<.001, η^2=.025), higher task value (F=12.14, p<.001, η^2=.021), higher academic self-efficacy (F=13.67, p<.001, η^2=.023), higher self-regulatory efficacy (F=37.60, p<.001, η^2=.062), and higher instructor evaluations (F=9.27, p<.001, η^2=.016) than the non-ELP group. Further, among the ELP pupils, those who used the portfolio more frequently (as reported by their teachers) showed significantly more positive self-regulated learning outcomes than those who used it less frequently. The authors conclude that "The European language portfolio accomplishes its pedagogical goal. Students in the experimental group using the ELP reported attributes more characteristic of self-regulated learners. […] The more frequently the European Language Portfolio is implemented, the better' (p. 933).

The evidence provided by this study is limited by the fact that the author provides no description of what implementing the ELP actually looks like, and does not take into account possible variation between schools or teachers. No information is provided concerning the language achievement data for the ELP and non-ELP cohorts, and of course there may also be all manner of other differences between the groups besides their usage or non-usage of the portfolio. For example, teachers who adopt the ELP may share particular characteristics or views of language learning – such as, potentially, a belief in the importance of developing metacognitively aware, self-regulating learners, thus leading them to choose to use the ELP in the first place. Finally, some may also question the fact that the statistical analyses performed in this study treat the five-point Likert scale measures as continuous variables. Once again, as a result of these limitations, we judged this study to have significant risk of bias. Nonetheless, it is again useful in raising teachers' awareness of the potential value of the European Language Portfolio, as well as of the importance of developing learners' self-regulation in general.

3.3.3 Conclusions from Harris and Ó Duibhir's (2011) and updated studies

Harris and Ó Duibhir's (2011) original review concludes that, "although there is a substantial body of research in the area of Second Language Acquisition, in many cases findings … are not sufficiently clear or uncontested enough to provide straightforward guidance for teachers" (p. 19). There are, they say, "no simple answers to the questions raised" in relation to policy and classroom practice. This is particularly the case once the enormous range of contextual differences between individual classrooms – and between the same classroom on different occasions – is taken into account. In light of this variety of contexts and constraints, it is unsurprising that research should be unable to provide 'off the shelf' solutions for individual teachers that they can apply irrespective of their particular class of learners at a particular time on a particular day. Indeed, in some ways the lack of 'simple answers' is reassuring: if these did exist, then surely we would all be confidently implementing them in classrooms around the world! In the words of Mitchell and Myles (1998:195), quoted by Harris and Ó Duibhir (pp. 69f):

> [M]ost importantly, teaching is an art as well as a science, and irreducibly so, because of the constantly varying nature of the classroom as a learning community. There can be 'no one best method', however much \ research evidence supports it, which applies at all times and in all situations, with every type of learner.

What this strongly implies is the preeminent importance of teachers' professional judgment in designing tasks to maximize learning outcomes in their own classrooms. We follow Winch et al. (2015) in arguing for an important role of research knowledge in shaping this professional judgment, but teachers also need to critically evaluate that research in light of their own particular teaching contexts. As Harris and Ó Duibhir argue, the value of reviews such as theirs (and the current REA) is that it provides an important source of knowledge on which teachers can draw to inform their decision-making in the classroom, rather than offering prescriptions for practice.

It is important, however, that such reviews gather the best available evidence. One of the striking features of Harris and Ó Duibhir's review is that many of the studies which result from their systematic search and the application of their inclusion criteria carry significant risk of bias as a result of multiple serious limitations. The same is true of the additional studies that we have located when updating their review. This begs the question of whether the search terms used are too restrictive. In seeking to obtain evidence which is as directly relevant as possible to classroom-based L2 teaching "in contexts similar to primary schools in Ireland" (p. 13), it seems that a good deal of other, more robust evidence has been excluded. The fact that this evidence may derive from, or relate to, rather different educational contexts does not necessarily mean that it cannot be relevant to teachers' decision making in the target context of the review. Once again, however, teachers' professional judgment comes to the fore here, since evidence derived from different contexts will need to be critically evaluated in order to gauge its relevance to a particular classroom.

3.3.4 Review questions 1 and 3: key conclusions

Notwithstanding the points we have raised about the importance of teacher expertise and contextual factors in determining the most effective approaches to teaching, in the table below we summarize what we consider to be the key conclusions and implications for RQs 1 and 3 of this rapid evidence assessment. These (listed in columns 1 and 2 of the table) arise from the Fitzpatrick et al (2018) and Harris and Ó Duibhir (2011) seed reviews, and our updates of them, and are those which we judge to be supported by the strongest evidence. In the third column, we then add our personal interpretation of these conclusions, drawing on our own experiences and expertise, and make specific recommendations for the teaching and learning of MFL in English primary and secondary schools.

Table 3.3 Overall conclusions from updates to Fitzpatrick et al (2018) and Harris and Ó'Duibhir (2011) – addressing review questions 1 and 3

Topic / focus	Implication	Additional comments	Studies Contributing to this conclusion
Oral corrective feedback	Corrective oral feedback can be effective in promoting L2 development, with prompts being more effective than recasts, which are in turn more effective than simply ignoring errors and providing no feedback.	We would specifically encourage this to be carried out within a context of high quality, spontaneous oral interaction, in which pupils try to formulate their ideas in the L2 beyond the use of preformulated chunks (Macaro, Graham & Woore, 2015).	Ammar & Spada (2006); Ammar (2008); Gattullo (2000).
Intensive L2 programmes	Intensive programmes of instruction in a second or additional language over a short time period are likely to be more effective than 'drip feed' programmes, where learners are exposed to limited amounts of the language over a longer	Intensive programmes may help learners make the "rapid and substantial progress" they need in order to sustain and develop their motivation for continuing to learn the language. However, this lies beyond the control of the individual classroom teacher and is more a matter for policymakers	Collins et al (1999), Collins & White (2011); Netten & Germain (2009).

	period of time. The nature of the language gains made will reflect the specific nature of the intensive practice that learners engage in.	and school leaders. It would require major system-level (or at least school-level) reform to make this possible in the UK.	
Communicative versus grammatical / analytical orientation of language programmes	Teachers should aim for a judicious balance of communicative and form-focused L2 instruction. However, what this balance looks like depends on the specific teaching context and so is a matter for teachers' professional judgment.	We fully endorse the importance of teachers' professional judgment and would advocate a research-engaged model of teacher professionalism (Winch, Oancea & Orchard, 2015), in which teachers draw on both their detailed contextual knowledge of their own classrooms and the available research evidence. Further, they should interrogate this research evidence critically in relation to their own particular context.	Alcón (2007); Arslanyilmaz (2013); Edelenbos & Suhre (1994); Griva & Semoglou (2012); Harris & Murtagh (1999); Ho & Binh (2014); Kasprowicz & Marsden (2018); Owen et al (2019);Van de Ven et al (2019).
Approaches to developing L2 grammar	Approaches and methods which incorporate speaking, reading and writing at primary level can be effective in developing grammatical knowledge. While explicit instruction in grammatical features (both deductive and inductive) is effective, it is rarely more effective than other types of instruction.	Approaches to teaching grammar should be tailored according to age and level of proficiency of the learners. Form-focused Instruction and Processing Instruction can be effective for teaching grammar, especially difficult concepts, but are likely to achieve better outcomes with KS3 and KS4 learners than younger learners. Both approaches are likely to be effective if there is contextualized input, the occurrence of negotiation of meaning, and student-initiated production.	Chan (2018); Graham et al (2017); Griva & Semoglou (2012); Hanan (2015); Ho & Binh (2014); Kasprowicz & Marsden (2018); Nutta et al (2002); Porter (2014).
Approaches to developing L2 vocabulary	While input-only instruction is effective in terms of (limited) vocabulary uptake, learning gains can be greatly enhanced when input is supplemented.	We would support supplementation to input in the form of pre-teaching and further interaction when attention and activity are focused on the form and meaning of individual vocabulary items. Vocabulary learning is also facilitated by tasks and games with a high involvement load. Teachers should carefully consider learners' proficiency level when adopting extensive or intensive reading approaches to vocabulary development.	Alcón (2007); Camo & Ballester (2015); Chen et al (2018); Chen, Liu & Todd (2018); Davis (2017); Dolean (2014); Dolean & Dolghi (2016); Gürkan (2019); Hennebry et al (2017); Huang et al (2012); Suárez & Gesa (2019); Laufer (2006); Laufer & Girsai (2008); Lee & Macaro (2013); Mavilidi et al (2015); Padial-Ruz et al (2019); Porter (2016),), Pujadas & Muñoz (2019); Sappathy (2011); Shintani (2013); Shintani (2012); Shintani (2011); Teng (2019b), Van de Ven, Segers, & Verhoeven (2019); Wang et al (2019); Williams & Thomas (2017).

Approaches to developing oral/aural skills	Strategy instruction and providing opportunities for students to develop strategies for speaking and listening are effective methods for improving both skills.	In relation to both speaking and listening, we would support the explicit teaching of strategies to help students overcome barriers to communication. Methods using engaging tasks and game-based activities that create meaningful interaction are more effective for the development of L2 speaking skills, pronunciation and communicative achievement than activities where the focus is exclusively on form. Such activities may be just as effective when conducted on-line as face-to-face. Technology such as automatic speech recognition software can be used effectively to develop learners' pronunciation and perception of foreign sounds; this is likely to be increasingly the case as technology continues to develop.	Cabrera & Martínez (2001), Graham & Macaro (2008); Griva & Semoglou (2012); Harris (2007); Jerotijević-Tišma (2016); Li et al (2017); Neri et al (2008); Nutta et al (2002); Rouhi et al (2014); Winasih et al (2019); Young & Wang (2014).
Approaches to developing L2 literacy skills	Primary MFL teaching should include a focus on developing literacy as well as oral/aural language skills.	At both primary and secondary levels, we would support the explicit teaching of both lower-level and higher-level reading processes, and reading skills and strategies, to help learners overcome difficulties they might encounter and gain access to a wider range of more interesting and challenging texts (see Woore et al., 2018). We would additionally advocate the teaching of phonics in the foreign language. Technology-supported reading can effectively contribute to the development of oral reading fluency. Effective approaches to teaching L2 writing require a wide variety of activities, including interactive writing tasks, and explicit teaching of strategies in order to take individual differences between learners into account.	Abdelrahman et al (2017); Amer (1997); Bartan (2017); Bruton (2007); Buga et al (2014); Chen et al (2018); Chen, Tan & Lo (2016); Coyle & Roca de Larios (2014); Drew (2009); Fidaoui et al (2010); Fonseca-Mora et al (2015); German et al (2005); Ghajar & Mirhosseini (2005); Gündüz & Ünal (2016); Gutiérrez Martinez & Ruiz de Zarobe (2017); Harris, 2007; Hwang et al (2014); Lan et al (2009); Lan et al (2015); Macaro & Erler, 2008; Macaro & Mutton, 2009; Manoli et al 2016; Mistar et al 2016; Porter (2014); Rostamian et al (2018); Sercu (2013); Takeda (2002); Tsiriotakis et al (2017); Teng (2019a); Türk & Erçetin (2014); Wang et al (2019); Yunus et al (2013); Zarobe & Zenotz (2015).
Teacher competence	Teacher language competence, experience and number of hours' instruction are more influential factors than instruction type.	Effective foreign language teaching is reliant on there being a sufficient supply of teachers with the relevant expertise, both in the languages themselves and in language pedagogy. It is therefore important that support be in place for teacher development and collaboration,	Graham et al (2017); Yunus et al (2013).

		within a model of teacher professionalism that recognizes the need for teachers to be able to use judgment and autonomy to respond to individual learning contexts.	
Computer-assisted Language Learning (CALL)	CALL can support school-based L2 learning, but the use of technology in itself is unlikely to bring benefits.	When technology and associated materials are well-targeted for specific activities and both teachers and learners are well trained in the use of various forms of technology, learning can be just as successful as with teacher-led activities; where activities are carefully planned, learners have control of the technology, and specific classroom management techniques are adopted, it can be even more successful and encourage learner autonomy, language awareness, self-regulation and engagement with authentic language. Learners with limited technological skills (and access to the technology itself) must be appropriately supported.	Abdelrahman et al (2017); Álvarez-Marinelli et al. (2016); Boulton & Cobb (2017); Buckingham & Alpaslan (2016); Buga et al (2014); Chen et al (2016); Fidaoui et al (2010); German et al (2005); Grgurović et al (2013); Gürkan (2019); Hwang et al (2014); Sercu (2013); Türk & Erçetin (2014); Wang et al (2019); Winasih et al (2019); Yunus et al (2013).
Video-supported L2 learning	The use of well-selected and graded L2 video clips and TV programmes can support L2 learning.	Especially with L2 captions and some guidance from teachers, video-supported activities can be an effective means of learning new vocabulary and developing listening and reading skills, as well as aural word recognition and decoding of foreign sounds, particularly where learners are able to control the pace of viewing in computer/language labs. Using peer-model videos with different planning foci can be motivating and effectively used to promote the accurate use of targeted grammatical structures and improve complexity during subsequent task performance.	Chen et al (2018); Collins & White (2011); Pujadas & Muñoz (2019); Teng (2019b); Teng (2019c); Suárez & Gesa (2019); Van de Guchte et al (2017); Williams & Thomas (2017).

3.4 Review Question 2

RQ2. What is the impact of learning a foreign language on students' wider academic outcomes?

The second review question of this REA asks whether there are any specific influences of foreign language learning on wider academic outcomes. Wider academic outcomes in this context was initially considered within educational provision where, for example, one might find research which suggests that learning a foreign language (FL) is associated with increased scores in mathematics. The results of our search as described in Section 2 (Methodology) illustrate that we were unable to find any seed reviews that specifically examined these questions within the context of academic content. Indeed, we would argue that this is a very important area for further study. We are aware of at least one randomised control trial which did investigate this question, however. In Murphy, Macaro, Alba and Cipolla (2015), native English-speaking students in year 3 (aged 7) were randomly assigned into one of three groups: i) Italian; ii) French and iii) Control. Children in the Italian group were taught Italian for one hour each week for a period of 15 weeks. Children in the French group were taught French for the same period. Children in the control group received no foreign language instruction[7]. All children were tested at pre- and post-test on their English (first language) reading and spelling skills. The main findings of Murphy et al (2015) illustrated that 15 weeks of FL instruction has a positive influence on aspects of L1 developing literacy skill – notably in phonological processing. Furthermore, children in the Italian group outperformed the French and Control groups on some measures suggesting that perhaps features of Italian (its transparent mapping of sounds on to phonemes, for example) might have a particularly facilitative influence on developing L1 literacy skills. This study addresses the focus of the second review question in that it illustrates that FL instruction can increase students' metalinguistic awareness, thus leading to increased outcomes on measures of developing L1 literacy. However, to our knowledge, there are very few studies of this type in the extant literature and as our search revealed, there are no systematic reviews or meta-analyses which have been carried out reviewing studies of this type.

[7] At the time this study was carried out foreign languages were not part of the primary curriculum in England.

Nonetheless, there has been a considerable amount of research which has examined the so-called advantages of being bilingual. Many researchers have for some time argued that bilingualism confers numerous advantages, ranging from the self-evident ability to communicate in more than one language, to the much-debated possibility that bilingualism leads to improved/enhanced executive functions. This research has tended to concentrate on relationships between bilingualism and cognitive functioning, but also include explorations of relationships between bilingualism and factors such as aging, communicative competence, metalinguistic awareness, intercultural competence, and creativity, among others. A tiny proportion of this literature focuses on the substantive effects of learning a foreign language on other academic related outcomes.

Of the reviews on these questions located in our original trawl, the highest quality and most recent was published in two parts: Fox, Corretjer, Webb and Tian (2019) and Fox, Corretjer and Webb (2019). These two reviews are essentially the same review but split into two where Fox, Corretjer, Webb and Tian (2019) (henceforth Fox et al.1) reviews research published between 2005-2011 and Fox, Corretjer and Webb (2019) (henceforth Fox et al.2) review research between 2012 and 2019. The focus of this two-part systematic review was to provide a comprehensive analysis of empirical research which speaks to the benefit of knowing more than one language. 'Knowing more than one language' is a somewhat vaguely specified phrase, but for Fox et al.1 and 2 this encompassed foreign language learning, bilingualism and multilingualism. As such we felt that including the two Fox et al reviews as seed reviews would therefore in part address the second review question of this REA as this work does speak to wider outcomes of knowing more than one language even if it is not always specifically concerned with academic subjects. They used a systematic approach to searching for literature, and performed both electronic and hand searches. Inclusion criteria are stated merely as "empirical research that was published in peer refereed journals" (Fox et al 2019:4). They identified 165 eligible studies, covering 19 different outcome domains, from executive functioning to attitudes toward other cultures. No trustworthiness appraisal is reported, aside from saying that they considered the limitations of each study. They did not report effect sizes or other statistical information from the findings of included studies, favouring a narrative synthesis. The nature of our Rapid Evidence Assessment was such that time did not allow us to locate all of the original studies cited by Fox et al., to assess their trustworthiness and extract the data necessary to calculate and report effect sizes ourselves. These data are not included in the report.

The two main research questions of the Fox et al.1 and 2 reviews were:

1. What are the effects of foreign language/world language learning and bilingualism on academic achievement, cognitive abilities, and learners' attitudes and belief?

2. What additional effects and factors may be associated with FL/WL learning or bilingualism drawn from the empirical research literature?

The method they adopted was based on Valdés, Kibler and Philipose (2004) who themselves were focused on three research questions: i) how language learning supports academic achievement; ii) how language learning provide cognitive benefits; and iii) how language learning affects attitudes and beliefs about language learning and other cultures. Fox et al. 1 and 2 (2019) used the same search terms as Valdés et al (2004) but they also added to these as they felt this would address what they refer to as 'emerging themes' from the original Valdés et al study. They followed the standard systematic review protocol where these terms were applied to a number of relevant databases to reveal a total of 65 studies which addressed some aspects of their review questions. They organised their presentation and discussion according to eight specific categories. These categories were: i) cognitive abilities and benefits; ii) academic achievement and benefits; iii) enhanced creativity; iv) communicative and intercultural competence; v) positive ageing; vi) positive attitude toward other languages/cultures; vii) greater earning potential and viii) motivation.

We updated Fox et al.'s work in line with the methods described previously (Section 2: Methodology), using the same search and inclusion strategy as the originals. We identified an additional 17 eligible studies. We assessed the trustworthiness of each study, per our protocol. Eleven studies were rated 2*, two studies were rated 1* and five studies were rated 0*. Where effect sizes were reported we extracted these and reported them as in the original reports. In reports where effect sizes were not reported, but in which sufficient data existed to allow us to calculate effect sizes ourselves, we did. In eight of the 17 studies sufficient data were not available.

Table 3.4: Update studies found from Fox et al 1 and Fox et al 2

Study	Topic	Context	Sample	Findings (including effect sizes, where given)	Trustworthiness rating in this REA
Aljohani (2016)	Effect of foreign language learning on L1 literacy.	Private and state primary schools in Jeddah, Saudi Arabia	2,000 primary school students.	Children at private schools, where English is taught, outperformed children at state schools (where English is not taught) on measures of Arabic literacy. No inferential statistics reported. No Effect Sizes reported. Insufficient data reported to calculate.	0
Anton, Carreiras & Duñabeitia (2019)	Executive functioning in adults.	Basque country	180 young adults. 90 balanced Basque-Spanish bilinguals. 90 Spanish monolinguals.	No statistically significant differences in outcomes between monolingual and bilinguals on flanker, Stroop, Simon, Coris, and digit span tasks. No Effect Sizes reported. Insufficient data reported to calculate.	2*
Comishen et al (2019)	Attentional control in infants.	Toronto, Canada	Six-month-old infants raised in monolingual (n = 10) or bilingual (n = 10) environments	Correct gaze direction following a visual stimulus statistically significantly different from chance among infants from bilingual households, but not among infants from monolingual households. Effect sizes: Correct anticipations: d = .48 Reactive latencies: d = .93	0
Damian, Ye, Oh & Yang (2019)	Executive functioning in young adults.	Bristol, UK. University context.	26 Chinese-English bilingual and 25 English monolingual undergraduate students.	When tested on Flanker, Simon and Stroop tasks, no statistically significant differences between monolinguals and bilinguals in response rate or accuracy. No Effect Sizes reported. Insufficient data reported to calculate.	2*
Fecher & Johnson (2018)	Communicative competence in infants.	Toronto, Canada.	48 8.5-9.5 month old infants. 24 monolinguals and 24 bilinguals	Statistically significant differences in gaze times suggested that bilingual infants noticed communicative incongruities more readily than monolingual infants. Effect size: Trial type x group (mono or bilingual): $\eta_p^2 = .08$	0
Festman & Schweitzer (2019)	Bilingualism and literacy skills.	Primary Schools, Berlin, Germany	125 third-grade children. 69 monolingual, 56 multilingual.	No statistically significant differences in performance on tests of reading comprehension and reading fluency between monolinguals and bilinguals. Monolinguals statistically significantly outperformed bilinguals on spelling tests. No Effect Sizes reported. Insufficient data reported to calculate.	1*
Gunzenhauser, Karbach & Saalbach (2019)	Impact of bilingualism on standardised test scores in children.	Elementary schools in Southwest Germany	3rd and 4th graders in Germany. Study 1: bilingual n=45, monolingual n=54	Monolingual children statistically significantly outperformed bilingual children on standardised tests of verbal competence. Effect sizes: Study 1: G = 1.16† Study 2 G = .96†	0

			Study 2: bilingual n=21, monolinguals n=36		
Kalia, Daneri & Wilbourn (2019)	Executive functioning in children.	Bilingual immersion and monolingual English elementary schools in USA.	61 Spanish-English and 55 monolingual Kindergarten – 3rd Graders.	Bilingual children statistically significantly outperformed monolingual children on executive functioning tasks. Effect sizes: *DCCS Task:* Correct Responses: η_p^2 = .; 09Perseverative errors: η_p^2 = .10; *LSS Task:* Correct Responses: η_p^2 = .05 Errors: η_p^2 = .05	2*
Lorge & Katsos (2018)	Communicative competence in adults	University students in Cambridge, UK.	40 adults aged 20 – 35, 10 monolingual, 10 bilingual.	Bilingual adults made statistically significantly more attempts at accommodating their interlocutors than monolinguals. No Effect Sizes reported. Insufficient data reported to calculate.	1*
Paap et al (2018)	Executive functioning in young adults.	University students, San Francisco, USA.	166 undergraduate students, 104 bilinguals and 62 monolinguals	No statistically significant differences between bilinguals and monolinguals on any of the measures of executive functioning. Effect sizes: *Response Time* - Group (mono or bi): η_p^2 = 0.003; Group x Congruency: $\eta_p^2 < 0.001$; Group × Task × Congruency: η_p^2 = 0.008; *Accuracy* -Group: η_p^2 = 0.004; Group × Task × Congruency: η_p^2 = 0.01; *Efficiency* - Group: η_p^2 = 0.001; Group × Congruency: $\eta_p^2 < 0.001$; Group × Task × Congruency: η_p^2 = 0.007	2*
Papageorgiou et al (2019)	Working memory and executive functioning in older adults.	UK	74 retirees, 37 lifelong bilinguals, 37 English monolinguals.	WM task - No statistically significant difference between monolinguals and bilinguals on digit span task (G = 0.09†). EF tasks - No statistically significant differences between monolinguals and bilinguals on: *Simon Task* Response Time: η_p^2 = .02; Congruency: η_p^2 = .01; Accuracy: η_p^2 = .97; *Tower of London Task:* Accuracy: G = .42†; *Change Blindness Task* - Response Time: G = .22† Accuracy: G = 10.54† ; Statistically significant differences, favouring monolinguals, on: *Tower of London Task*- Response Time first move: G =.44†; Response Time trial completion: G = .43†	2*
Paplikar et al. (2019)	Bilingualism and aphasia following stroke.	Hyderabad, India.	65 ischemic stroke patients, 27 monolingual, 38 bilingual.	Aphasia following stroke was statistically significantly less severe among bilinguals. Effect size: ACE-R: d = −0.691	2*

Robinson & Sorace (2018)	Executive functioning in young children.	Socially disadvantaged areas in Edinburgh, UK.	62 children aged 5 to 6 years, 36 English monolinguals, 26 balanced and unbalanced bilinguals.	Bilingual children statistically significantly outperformed monolinguals on DCCS task. Authors emphasise potential for environmental factors to explain differences. No effect sizes reported. Insufficient data reported to calculate.	2*
Singh et al. (2019)	Attitudes towards other ethnicities in infants.	Singapore	72 infants aged 18-20 months. 36 from bilingual households, 36 from monolingual households.	Statistically significant difference in gaze characteristics suggested greater trust in adults of different ethnicity among infants from bilingual households compared to infants from monolingual households. No Effect Sizes reported. Insufficient data reported to calculate.	2*
Tran, Arredondo & Yoshida (2019)	Development of executive functioning in young children.	USA, Vietnam, and Argentina	96 3-year-olds. 44 bilinguals, 49 monolinguals.	Bilingual children statistically significantly outperformed monolinguals on DCCS, Day/Night, and Gift/Delay tasks. No statistically significant difference on the Bear/Dragon task. Effect sizes: DCCS task: 0.34; Day/Night task: 0.21; Gift/Delay task: 0.12Bear/Dragon task: 0.07; (Authors do not state which statistic was calculated; D, G or η_p^2)	2*
Valis et al. (2019)	Effect of learning a foreign language on cognitive functioning (memory and attention).	Adult education, Chechia.	42 cognitively unimpaired Czech citizens, mean age 70.9 years.	12-week English language course was not associated with statistically significant differences between bilinguals and monolinguals in memory and attention. No Effect Sizes reported. Insufficient data reported to calculate.	1*
van Veen et al. (2019)	Cognitive development in infants born pre-term.	Dutch primary schools.	234 monolingual, 91 bilingual children. All born pre-term.	Monolingual infants outperformed bilingual infants on measures of cognitive development, at both 2 and 5 years. Effect sizes: Bayley-III at 2 years: d = .42; WPSSI-III-NL at 5 years: d = .48; Verbal IQ at 5 years: d = .65 General language composite at 5 years: d = .55	1*

We have structured our discussion of this work around the same themes as the original Fox et al 1 and 2 reviews. In so doing we both describe the content of the original Fox et al seed review and then present updated research we found where relevant within each section.

3.4.1 Cognitive abilities

By far the most heavily researched area on the potential benefits associated with knowing another language (beyond the obvious: knowing another language) is the relationship between bilingualism and cognitive abilities. This type of research is characterised in the main by selecting two groups of individuals, one of which is bilingual and the other of which is monolingual, and comparing their performances on tests of cognitive ability. These tests usually consist of presenting a series of stimuli to the participants and asking them to respond. For example, in a test called the Simon task, participants are shown images of either triangles or squares on a computer screen. They are told to press a button to the left of the screen when they see a triangle and to press a button to the right of the screen when they see a square. Sometimes the positioning of the stimulus is congruous with the expected response where, for example, a triangle (which requires a 'left' response) appears on the left of the screen. Sometimes the placement is incongruous with the expected response such as when a triangle is placed on the right of the screen, but still requires a 'left' response. Because incongruity requires more cognitive processing than congruity, and therefore more processing time, researchers use the time it takes from the presentation of the stimulus to the pressing of the correct button as a measure of cognitive functioning. Response times can then be compared across groups to assess whether knowing another language is associated with a difference in response times, and therefore cognitive functioning.

Other tasks used to assess cognitive functioning include Flanker tasks, where respondents must process incongruity between a stimulus and items placed either side of it; Stroop tasks, where a colour word is written in ink that is either congruous or incongruous with the word (for example, the word 'red' is written either in red ink (congruous) or in blue ink (incongruous)), and the respondent has to name the colour of the ink; and many others. Other tests are used to measure problem solving (such as the Tower of London task, where participants must work out how to transfer stacked coloured rings from one tower to another tower to end up with a specific order of colours); verbal working memory (for example the digit span task, where participants are asked to remember and repeat sequences of numbers presented in increasing length); and visuo-spatial working memory (for example the Change Blindness task, where two almost identical images are presented in rapid repeating succession and the participant must identify what is different about the images as quickly as possible).

The theoretical argument as to why there might be differences in performance on these tasks between bilinguals and monolinguals relates to the constant cognitive control required when more than one language is present in the mind. We understand that all languages known to an individual are always activated (Green 1999). Therefore, suppression of one language is required when communicating in another. The near constant practice of managing languages in this way is regarded much like a workout for the physical muscles. Someone who spends time regularly exercising muscles by lifting weights in a gym is at an advantage when required to use those muscles for other tasks, like opening jam jar lids. By the same token, someone who spends time regularly exercising cognitive 'muscles' through controlling their languages is thought to be at an advantage when cognitive control is required for other tasks, such as solving problems or ignoring irrelevant information while concentrating on relevant information.

There are three issues with this research that we would like to draw attention to before summarising the evidence collected by Fox et al. and the additional evidence that we have located. The first is that true experimental conditions are very difficult to contrive in this research. Often in these experiments, groups of people who are already bilingual are compared to those who are not. This means that comparison groups are systematically different at baseline. As such we cannot disentangle the effects of bilingualism from other possible effects of belonging to a bilingual community, and thus clear casual relationships cannot be asserted. The second is that response times and other measures of cognitive control are proxy measures for the things that we are actually interested in. A ten-millisecond advantage in responding to a Simon task is only interesting to teachers and educational policy makers if it translates into improved performance on real world tasks, such as solving maths problems or managing complex information. Unfortunately, research in this area rarely goes the extra step to assess whether having learned another language extends beyond test performance to translate into improved performance on real world tasks. The third issue relates to publication bias and questionable reporting practices associated with this kind of research.

Publication bias, where results of studies that contradict the orthodoxy of the field are either not forwarded for publication by authors or are not accepted for publication by journals, is certainly not restricted to the literature on bilingualism. But in other fields it is often suspected, rather than confirmed. Recent scholarship in bilingualism research has provided empirical evidence that publication bias is almost certainly at play in this field (de Bruin et al. 2015; Lehtonen et al. 2018). These researchers found good evidence that research on bilingualism and cognitive function is much more likely to be published if it confirms a so-called bilingual advantage, and much less likely to be published if it suggests no difference between bilinguals and monolinguals, or if the advantage is found in monolinguals. They also found evidence of selective outcome reporting. That is, researchers tend to test bilinguals and monolinguals on a wide range of different tasks such as those described above, but only report the outcomes of tasks where a bilingual advantage was detected. This is akin to flipping a coin ten times, it coming down tails on the first eight flips and heads on the final two, then presenting only the results of the final two flips to argue that the coin is double-headed. Both of these practices provide a skewed representation of research in this area and therefore make it very difficult to have confidence in its collective findings.

Finally, as indicated in Section 2 (Methodology) we applied Gorard's (2014) sieve to examine quality/weight of evidence (see Table 3.4 and Appendix 11 for individual ratings). However, given the majority of the studies in this section fall outside of the parameters of an educational intervention (i.e., not an RCT or QED that is also an intervention), some of the criteria of the sieve itself (e.g., attrition, fidelity to treatment) were not relevant and consequently, this assessment of weighing the evidence is less than ideal, albeit consistent with the rest of the research in this REA. The highest overall rating provided in Table 3.4 and Appendix 11 was 2* despite the fact that for many studies, the individual parameters of Gorard's sieve were ranked much higher. Indeed, most of the studies (with only 3 exceptions) consistently received rankings of 4 on features such as scale, dropout, outcomes and validity, but nonetheless received low rankings due to design. This is largely attributable to the fact that many studies in this review adhered to a matched comparison design where bilinguals matched against a group of monolinguals were recruited and compared on some measure. Fidelity and Dropout are two further criteria of Gorard's sieve that do not easily apply to most of the research in the Fox et al.1 and 2 seed reviews and update. The lowest rating for Outcomes was 3, and the ratings ranged from 2-4 for Scale. As such, many studies are ranked much lower than the quality of the work in general should illustrate, in our view. As stated in our Methodology (Section 2) Gorard's sieve is a very strict instrument and one that does not adequately capture the variety of experimental designs within bilingualism research. Hence a low rating on this measure is not necessarily indicative of poor-quality research. Nonetheless, as with all research, the evidence in this domain of enquiry is mixed in terms of methodological rigour where some studies were allocated overall ratings of 0 (Comishen et al (2019); Fecher & Johnson, (2018); Gunzenhauser et al (2019)) or 1 (Valis et al (2019); van Veen et al (2019)).

Research on the so-called bilingual advantage has enjoyed much publicity, has been a feature in the popular press about the benefits of language learning, and has shaped thinking about how we explore these relationships. It would, therefore, be remiss of us not to include it in this rapid evidence assessment. However, we do so with a strong warning that serious methodological questions persist over findings in this area, and it should be interpreted with this in mind. Caveat lector.

3.4.1.1 Cognitive control

Executive functioning is defined by Kroll and Dussais (2017) as the functions that allow humans to execute complex tasks. This includes control of attention, inhibition of distractions, working memory, self-monitoring and switching between tasks. Fox et al. located a number of relevant studies which will be reported by theme in what follows.

Executive functioning

Executive functioning is the term used to denote the cognitive functions that allow humans to engage in complex tasks (Kroll & Dussais, 2017). These functions typically include: controlled attention, inhibition, distraction, monitoring, working memory and task switching. Fox et al. 1 identified 12 studies which indicated the bilingualism enhances executive functioning. For example, Bialystok (2007) and Festman, Rodriguez-Fornells and Muente (2010) and Hernandez, Costa, Fuentes, Vivas and Sebastian-Galles (2010) all suggest that controlling attention of two languages boosts executive control processes in childhood bilinguals. Vega and Fernandez (2011) provide some evidence to suggest that these enhanced control processes are moderated somewhat by the degree of balance within the bilingual – i.e., the extent to which the child has equal skills in both of their languages. Being able to suppress task-irrelevant information has also been argued to be superior in bilinguals (e.g., Blumenfeld & Marian, 2011; Carlson & Meltzoff, 2008; Colzato et al., 2008). Poulin-

Dubois, Blaye, Coutya and Bialystok (2011) specifically claim that toddlers (24 months old) were significantly better than monolingual toddlers on a Stroop task (where the participant has to ignore competing information to complete the task accurately). This is one of the first (and only) studies to demonstrate this finding in participants this young.

Fox et al. 2 (2019) identified 19 studies which claimed to demonstrate heightened executive functioning. However, in this part of the review they also identified studies which questioned this advantage of bilingualism on executive functioning. Abdelgafar and Moawad (2015) and Ross and Melinger (2017) produced equivocal findings regarding bilingual children's performance on executive function tasks. This lack of direct evidence for a bilingual advantage of executive functions was also manifest in von Bastian, Souza and Gade (2016) Karlsson et al (2015), and Ross and Melinger (2017). The mixed evidence has been argued either to indicate a complex relationship with age (where age moderates the impact of bilingualism on executive functions such as in Costa, Hernández, Costa-Faidella and Sebastián-Gallés (2009), language environment (Kousaie, Sheppard, Lemieux, Monetta & Taler, 2014), or, that there is just simply no credible evidence to support this claim (Paap, Johnson & Sawi, 2014; Papageorgiou, Bright, Tomas & Filippi, 2019).

In our update to Fox et al. 1 and 2 (2019) we found an additional eight studies which addressed the question of a bilingual advantage on executive functions. Antón, Carreiras and Duñabeitia (2019) tested young adult bilinguals and compared them against monolingual controls on a range of tasks used in the past literature on executive functions (e.g., Flanker, Simon, Stroop). Their findings were again equivocal and they argue that the putative bilingual advantage on executive functioning is not due to bilingualism per se but rather, other confounding factors which were/have not been addressed in studies purporting to demonstrate the advantage.

In Damian, Ye, Oh and Yang (2019) English monolingual participants were compared to Chinese-English bilingual young adults on the usual tasks (Flanker, Simon, and Stroop). A key difference in their study was that instead of asking the participants to use the standard key press to indicate their response, participants use a mouse which they suggested might be more sensitive. Their data suggest that bilinguals produced more 'efficient' responses than monolinguals but equally they are reluctant to argue for a bilingual advantage on executive functions, mostly because bilinguals and monolinguals in general performed identically in terms of accuracy and response rate.

Kalia, Daneri and Wilbourn (2019) tested dual language (Spanish-English) children in the US on measures of vocabulary and two measures which they argue tap in to executive functions: i) the dimensional change card sort task (DCCS) where children have to sort cards based on both shape and colour; and the ii) lexical Stroop sort (LSS) task where children have to pay attention to words when conflicting with colour (or objects). As is common in studies of dual language learners, monolingual children had higher scores on the standardised (English) vocabulary test than the Spanish-English bilingual children. However, despite having lower English vocabulary scores the bilingual children were more accurate on the DCCS and LSS tasks in comparison to the monolingual children. Kalia et al (2019) suggest that while their results do suggest some executive function advantages for bilingual children, they also note the significant debate in the literature and the possibility that other variables (such as use of metacognitive strategies) might be the contributing factor to their results rather than bilingualism per se.

Paap, Anders-Jefferson, Mikulinsky, Masuda and Mason (2019) tested 104 bilinguals and 62 monolinguals on the Simon, vertical Stroop, spatial Stroop and Flanker tasks – all of which are examples of nonverbal interference tasks and are widely used in studies of this nature. They found no group differences on any of the tasks and consequently failed to find a bilingual advantage on these dimensions of cognitive function. This lack of bilingual advantage was also the outcome of Papageorgiou, Bright, Tomas and Filippi's (2019) study comparing bilingual and monolingual adults who were matched on age, gender and SES. As with Paap et al (2019) no group differences were found and they argue there is little credible evidence for a bilingual advantage on executive functions from their research.

Robinson and Sorace (2019) were more positive in their assertion that bilingual children have cognitive (executive function) advantages. In their study, five-six year old children were recruited from public primary schools in a disadvantaged area of Edinburgh. There were 26 bilingual (unbalanced) children and 36 English monolingual children. As with Kalia et al

(2019), children were administered the DCCS (among other measures), and similarly found bilingual children performed better on the DCCS than the monolinguals.

Tran, Arredondo, and Yoshida (2019) tested even younger children in their study. They recruited bilingual and monolingual children between starting at age three and investigated how the development of early executive function skills might develop in these children. Four different executive function (EF) tasks were administered at three time points: when the child was 3, 3.5 and then at 4 years. They found advantages for the bilingual children on the DCCS and two of the other EF tasks. In addition to arguing for a bilingual advantage, they also argue that cultural aspects influence performance on these tasks.

The final study we found in our update of Fox et al. 1 and 2 (2019) which examined executive functioning in bilinguals was Valis, Slaninova, Prazak, Poulova, Kacetl and Klimova (2019). In this study, foreign language (not bilingual) learning was investigated with respect to cognitive functions. They recruited 42 Czech participants who were randomly assigned to either an English foreign language learning group (where they learned English for 12 weeks), or a control group. They measured cognitive functioning somewhat differently from the other studies described here, through the administration of the Montréal Cognitive Assessment which is argued to tap in to memory and attention processes which form part of executive functions. The post-test scores did not demonstrate any differences between the two groups and consequently the authors reject the hypothesis that cognitive functions were enhanced through foreign language learning in their study.

In summary, our update studies are in line with Fox et al. 1 and 2 (2019) original review in that at best mixed evidence is found to support the claim that learning a foreign language, or being bilingual can enhance executive functions. Whereas some studies were clear in arguing for such an advantages, others were equally clear in suggesting there is no such advantage.

3.4.1.2 Inhibitory control

Fox et al. found several studies that assessed inhibitory control in bilinguals, most of which found advantages for bilinguals. Studies with pre-adolescent Spanish-English bilinguals and

English monolinguals (Park, Ellis Weismer, and Kaushanskaya 2018), with young adult Kurdish-Persian bilinguals and Persian monolinguals (Kazemeini and Fadardi 2016), and with adult Spanish-English bilinguals and English monolinguals (Fernández, Tartar, Padron, and Acosta 2013) all found that the bilingual groups demonstrated significantly better inhibitory control than monolinguals. These basic findings were also reported in studies with adults, adolescents, children and toddlers (Blumenfeld & Marian, 2011; Carlson & Meltzoff, 2008; Colzato et al. 2008; Poulin-Dubois, Blaye, Coutya, & Bialystok 2011; Soveri, Laine, Hamalainen, & Hugdahl 2011; Stafford 2011; Wodniecka, Craik, Lin, & Bialystok, 2010).

Choi, Jeon, and Lippard (2018) explored inhibitory control in kindergarteners who were defined as either Spanish-English balanced bilinguals, Spanish-English partial bilinguals, and English monolinguals. They found that the balanced bilinguals outperformed both other groups on tests of inhibitory control. A study of Walloon kindergarteners (Woumans, Ameloot, Keuleers, & Van Assche 2019), cited by Fox et al. as supporting the idea that L2 learning improves inhibitory control, on closer inspection turns out to be evidence that pre-existing inhibitory control is predictive of L2 learning. In this longitudinal study, monolingual French children were tested for inhibitory control on entry to a Dutch immersion kindergarten. After one year, all children had improved in their Dutch proficiency and their inhibitory control. However, there was a strong predictive relationship between inhibitory control at baseline and Dutch proficiency after one year. This speaks directly to one of the main points of uncertainty over the methodological character of much of this research; that of causal direction. By way of illustration of how this uncertainty manifests, Sun, Li, Ding, Wang, and Li (2019) compared the extent of inhibitory control among young adult Mandarin-English bilinguals classified as either high proficiency or low proficiency. They found that high proficiency bilinguals have better inhibitory control than low proficiency bilinguals. They leave the direction of this relationship open to interpretation.

3.4.1.3 Working memory

All studies included in the Fox et al. review (Marini, Eliseeva, & Fabbro 2019; Cockcroft, Wigdorowitz, & Liversage 2019; Blom, Küntay, Messer, Verhagen, & Leseman, 2014; Macnamara & Conway 2014; Schroeder & Marian, 2012; Jiao, Lui, Wang, & Chen 2019; Babcock and Vallesi 2017) asserted support for a positive relationship between bilingualism and working memory, but especially where exposure to another language was early and sustained (i.e. from birth/early childhood and into schooling). They cite Marini et al. (2019) as stating that working memory is still developing during this period and thus suggest that early L2 learning shapes this process.

We located one additional study that examined working memory and which met the Fox et al. inclusion criteria. Papageorgiou et al. (2019) explored working memory with adults in the UK who had been everyday users of more than one language for more than 50 years. Their working memory was compared with that of similarly aged monolinguals. No difference was detected.

3.4.1.4 Attentional control

Attentional control is the capacity of an individual to choose what they pay attention to and what they ignore (Astle & Scerif 2009). Fox et al. found five studies that included exploration of attentional control. Two studies (Blom, Boerma, Bosma, Cornips, & Everaert, 2017; Chung-Fat-Yim, Sorge, & Bialystok, 2017) looked at bilingualism and selective attention. Two studies explored attention in disadvantaged young bilingual learners (Ladas, Carrol, & Vivas, 2015; Yang & Yang, 2016), and one study on attentional control in older adults (Ong, Sewell, Weekes, McKague, & Abutalebi, 2017). They all found that bilinguals had better attentional control than monolinguals, but tended to stress that this was associated with lifelong bilingualism, starting before school and persisting into late adulthood. The implications for foreign language teaching are therefore unclear. In the study by Ladas, Carrol and Vivas (2015), there is evidence that among disadvantaged young bilinguals (6-8 year old Albanian-Greek bilingual immigrants) there is no advantage of their bilingualism on attention. This provides a useful insight into the strength of purported bilingual advantages relative to other characteristics of learners. That is, any advantage that exists may be negligible in relation to other influences on children's educational development.

We found one additional study on attentional control that met Fox et al.'s inclusion criteria. Comishen et al. (2019) examined the selective attention of six-month old infants, ten from bilingual households and ten from monolingual households. They did this by showing them stimuli that were preceded by a cue. To begin with the cue was reliable, it accurately predicted where the stimulus would appear. Later it switched to being unreliable relative to where the stimulus subsequently appeared. The eye movements of the infants were tracked to examine whether they updated their expectations of where the stimuli would occur after the switch. Comishen et al. report that "only infants raised in bilingual environments successfully updated their expectations" (2019:8) following the switch. They conclude that bilingual environments contribute to the development of better attentional control.

3.4.1.5 Cognitive flexibility

The definition of cognitive flexibility used by Fox et al. is somewhat broad; "the ability to shift between mental sets, in cognitive tasks and social interactions" (2019:708) and they report on five studies to assert a relationship between bilingualism and cognitive flexibility. The first (Kuipers & Thierry 2013) used event related brain potential pupil size in bilingual and monolingual toddlers to argue that bilinguals have better cognitive flexibility. They did this by measuring pupil dilation. When they were shown unrelated pictures, the bilingual toddlers' showed a greater pupil dilation response than the monolinguals. This, they argue, is evidence for better attentional response and greater flexibility, which in turn suggests a superior ability to shift between mental states. The other study (Marzecová et al. 2013) measured attentional shifts in 22 Hungarian monolingual and 22 Hungarian-English adults. They found that bilinguals showed reduced switch costs compared to monolinguals, and thus greater cognitive flexibility. A study of bimodal bilingualism (signing and writing) among deaf American adults (Kushalnagar, Hannay, & Hernández, 2010), a study of task switching in bilingual and monolingual American college students (Prior & MacWhinney, 2010) and a study in which Arabic-Hebrew bilingual children and their monolingual peers drew pictures of imaginary houses and flowers (Adi-Japha, Berberich-Artzi & Libnawi 2010) all assert a bilingual advantage in cognitive flexibility. They all emphasise the theory that the frequency with which bilinguals routinely switch languages provides cognitive training that is then transferable to other forms of cognitive switching.

3.4.2 Linguistic processing and reasoning

The definition of linguistic processing and reasoning given by Fox et al. is "how the use of words to communicate ideas, feelings, and communications is understood" (2019a:477). Siegal et al. (2010) explored the effect of second language learning among young German-

Italian bilinguals and English-Japanese bilinguals on conversational understanding. Children in Italy whose first language was German were found to statistically significantly out-perform Italian monolinguals on their handling of conversational maxims. That is, to be informative and avoid redundancy, speak the truth, be relevant, and be polite (Grice 1975). English-Japanese bilinguals in England had statistically significantly better vocabulary than Japanese monolinguals in Japan. These findings were used to argue that bilingualism contributes to better conversational competence. Kaushanskaya, Yoo and Marian (2011) studied the influence of foreign language learning on first language vocabulary and reading among English-Spanish and English-Mandarin bilingual adults. They found mixed results suggesting that second foreign language proficiency can either enhance or diminish first language competence in these areas. The data are all from self-reported perceptions about proficiency. The divide existed along language lines; the English-Spanish bilinguals reporting positive relationships, the English-Mandarin bilinguals reporting negative relationships. The authors suggest that linguistic distance between languages might explain these different findings.

A dataset analysis of established bilingual Spanish-English children, emerging Spanish-English bilingual children and monolingual English children by Choi, Rouse, and Ryu (2018) showed that bilingual and monolingual children similar developmental trajectories in skills associated with linguistic processing and reasoning. However emergent bilinguals consistently underperformed relative to the other groups, specifically in their vocabulary development.

In a study of Brazilian adults Thompson (2013) found that language aptitude was significantly related to experience with other languages. Fox et al. suggest that this reflects the positive influence that additional language learning can have on individual's ability to process and reason. They further this argument by citing the results of a comparison of monolingual and bilingual children's resolution of referential conflict by Verhagen, Grassmann, and Küntay (2017). Here two to four-year-old Dutch-English bilinguals and Dutch monolinguals were presented with a task that contrived conflict between pointing and labelling. The experimenters asked the children to take an object. However, the object that they named was not the object that they pointed to. Bilingual children were more likely to take the object pointed to than the object named. The authors argue that this represented a difference in the way that bilingual and monolingual children process pragmatic cues, suggesting that bilinguals are more sensitive to them then monolinguals. Marinova-Todd (2012) supports this general view. In her study of 20 monolingual and 20 bilingual third-grade children in the USA, she found that bilinguals were better than monolinguals at establishing word meaning from context.

In a study that emphasises the need to take a longer view of the effects of foreign language instruction, Jaekel, Schurig, Florian, and Ritter (2017) compared linguistic development in German children who had been learning English from age six with those who had been learning since learning English at age eight. When tested on picture recognition, sentence completion, and reading comprehension at age ten to eleven, the early start group outperformed the later start group. However, when they were re-tested at age twelve to thirteen, the later start group had overtaken the early start group and were statistically significantly out-performing the early start group on reading and listening. Despite folk wisdom suggesting that earlier is better for learning another language, this study provides evidence that this is not necessarily the case (a finding that has been long understood to be the correct interpretation, see Lightbown 2008). The theoretical argument for this is that older learners have better established proficiency in linguistic processing in their first language than younger learners, which they can then apply to their new language learning.

3.4.3 Metalinguistic awareness

Meta-linguistic awareness refers to the capacity to talk about, analyse, and think about language independently of literal meanings. Meta-linguistic awareness is an understanding of how a language works. In studies of language learning, it is thought that knowing more than one language contributes to meta-linguistic awareness because this knowledge reveals similarities, differences and ultimately the arbitrariness of languages (it’s an apple in English, a pomme in French; the adjective precedes the noun in English, but comes after it in French, and so on). Fox et al. found 15 studies that addressed the relationship between bilingualism/knowing another language and meta-linguistic awareness.

Thirteen of these were studies with children (Dillon 2009; Ibrahim, Eviatar, & Aharon-Peretz 2007; Kuo & Anderson 2010; Schoenpflug & Klische 2010; Verhoeven 2007; Bialystok & Barac, 2012; Diaz and Farrar’s 2018; Bien-Miller, Akbulut, Wildemann, and Reich 2017; Daller and Ongun's 2018; Hermanto, Moreno, & Bialystok, 2012; Reder, Marec-Breton, Gombert, and Demont 2013; Sun, Hu, and Curdt-Christiansen 2018; Yeon, Bae, and Joshi 2017). All converged on the same essential finding: that bilingualism is associated with better meta-linguistic awareness compared with monolingualism. This was true of phonological awareness, phonotactic awareness, morphological awareness, and syntactic awareness. In particular, the importance of a strong, regularly used home language was reported frequently as being a strong predictor of meta-linguistic awareness among bilingual children. While most studies addressed differences between already bilingual children and monolinguals, the study by Reder, Marec-Breton, Gombert, and Demont (2013) investigated children in France learning German. This study, therefore, may have more relevance to pedagogical policy for children who do not come from bilingual households. Forty-three first graders who had been learning German since kindergarten (aged 4) in a partial immersion programme (where half of the instructional time was in French and the other half was in German) were compared to children who had come up through a French only kindergarten. They found that the German learners had superior scores on tests of compounds morphological awareness and syntactic awareness, but not of phonological awareness or affixes morphological awareness. Similarly, a study by Laurent and Martinot (2010) demonstrated that children who have been learning another language through bilingual programmes of education, start to show superior phonological awareness compared to their monolingual counterparts by age nine. This difference strengthens over time.

Two studies in the Fox et al. literature investigated adult learners (Huang 2018; Ransdell, Barbier, & Niit 2006. Both found advantages among adults who know another language and suggest the importance of developing this kind of knowledge for learning a third language.

We did not find any additional studies that addressed language learning/bilingualism and meta-linguistics awareness that met Fox et al.'s inclusion criteria.

3.4.4 Cognitive development

Fox et al. define cognitive development as "the construction of thought processes such as remembering, problem solving, and decision making, from childhood through adolescence to adulthood." (2019a:478). They report on two reviews of the literature. (Bialystok & Craik 2010; Lazaruk 2007). Bialystok and Craik (2010) reviewed literature, which in the main focused on lab studies and found that bilingualism is associated with lower formal language proficiency than monolinguals, evidenced through smaller vocabularies and weaker access to lexical items, but better executive control in nonverbal tasks requiring conflict resolution. Lazaruk's (2007) review explored the effects of teaching French to monolingual English speakers in Canada, through immersion models of education. They found that children taught French in this way had superior outcomes in cognitive development and academic attainment.

They also report on primary studies of brain imaging (Arredondo, Hu, Satterfield & Kovelman 2017; Bartolotti, Bradley, Hernandez & Marian 2017; Jasinska & Petitto 2013; Kuhl et al. 2016) all of which noted structural differences in the brains of bilinguals compared to monolinguals, in what Fox et al. call the "classic language areas of the brain" (2019b:710). They suggest that this reflects important influences of early and sustained bilingualism on cognitive development.

In a study of low socio-economic status Spanish-English pre-schoolers in Head Start programmes in the USA, Santillán and Khurana's (2018) results suggested that facility with another language and sustained use of more than one language was associated with rapid development during the transition from preschool to kindergarten.

Fox et al. note that new research in this area (e.g. Folke, Ouzia, Bright, De Martino & Filippi 2016; Struys, Duyck & Woumans 2018) are starting to question the orthodoxy in this area, and have noted that differences between bilinguals and monolinguals might be better explained by the strategic choices these two groups make, rather than by a difference in cognitive architecture.

We found one additional study that addressed cognitive development in bilingual children. van Veen et al. (2019) conducted a retrospective cohort study on children who had been born prematurely, at ages two and five years. They compared 65 children from bilingual households with 234 monolingual children on standardized tests of cognitive development (Bayley-III at two years and WPPSI-III at five years), adjusted for premature birth. They found that monolinguals statistically significantly out-performed bilinguals at both time points.

3.4.5 Cross-language activation

Cross-language activation refers to the idea that a bilingual has access to both languages, and both are always activated to some degree. Language switching, where the bilingual has to switch seamlessly between their two languages, is often used as a measure in examining cross language activation. In Fox et al 1 (2019) Costa, Santesteban and Ivanova (2006) demonstrated that high proficiency Spanish-Basque and Spanish-English bilinguals manifest symmetrical switching across the two languages, regardless of the age at which they began learning their second language. Kormi-Nouri et al (2008) recruited a large sample of Swedish-Persian bilingual and Swedish monolingual children and argued for a positive impact of bilingualism on switching tasks. Of course, language switching within a monolingual is an impossibility, nonetheless language switching is associated with executive functions (EF) and could be argued to promote enhanced EF in bilinguals.

3.4.6 Spatial reasoning

Two studies in Fox et al.'s review addressed bilingualism and spatial reasoning. Greenberg, Bellana, and Bialystok (2013) investigated bilingual and monolingual eight-year-olds' perspective taking. Participants were shown an array of blocks and asked to decide how that array was seen by another observer from a different perspective. Bilingual participants were better at identifying the correct orientation of the blocks. The authors argue that this may be evidence for better academic attainment in bilinguals, citing the relationship, observed elsewhere, between spatial reasoning and IQ. Stephens and Moxham (2019) assessed spatial reasoning among 173 medical students in Cardiff. They used an online spatial awareness questionnaire and information about the students' linguistic backgrounds and found that bilingual students performed better on the task than monolinguals. The difference was more pronounced in students whose other language was non-European. The motivation for the study was related to performance on anatomy exams, which requires spatial awareness relative to the human body. While bilinguals demonstrated better spatial awareness in the questionnaire, this did not translate into better performance on the exam.

3.4.7 Academic achievement

Fox et al.'s findings on research that assess relationships between learning a foreign language/bilingualism and academic achievement can be divided into three themes:

language and literacy skills, academic attainment, and impact on standardized test scores. In interpreting this literature, it is important to understand the three fundamental aetiologies that tend to form the focuses of this research: being bilingual, attending a bilingual school, and studying a foreign language. 'Being bilingual' is somewhat of a trait rather than a state. That is, many of these studies compare people who are already bilingual, who come from bilingual homes and communities, and who have made use of more than one language since birth or very early childhood with people who are monolingual. There are many characteristics associated with people who are bilingual, aside from their bilingualism, that offer competing explanations for any differences in achievement between them and monolingual people. As such, inferring causal relationships and mechanisms is complicated. 'Attending a bilingual school', also carries a causal inference warning. Children (or parents) who choose bilingual programmes over monolingual programmes are, by definition, systematically different to their peers who choose monolingual programmes. Studies in this field frequently do not assign participants on the basis of chance to bilingual or monolingual programmes, instead analysing data from children already in them (but see Steele et al. (2017) for a rare exception to this general rule). Thus, research on bilingual education must be considered in the light of this methodological idiosyncrasy. Explorations of the effects of 'studying a foreign language' are potentially more informative to the purpose of this REA, as they explore the effects of deliberate teaching of foreign languages on academic outcomes. Theoretically at least, studies of this sort can more easily adopt research designs that control for systematic variation among participants, such as using unbiased methods to allocate participants to conditions.

Fox et al.'s findings in each of these areas are presented below followed in each section with any new studies we located in our update.

3.4.7.1 Language and literacy skills

Comparisons of bilingual and monolingual people, with no direct pedagogical component, have asserted better responses among bilinguals to early literacy teaching (Silven & Rubinov's 2010), better performance on the reading comprehension component of the Cambridge First Certificate of English (Modirkhamene, 2006), better word identification and oral language skills (Knell et al., 2007), and better vocabulary recall (Bialystok & Feng, 2009) than monolinguals.

In studies of bilingual education programmes, instruction in more than one language has been associated with higher levels of linguistic performance in both languages, at no cost to academic performance (Lazaruk, 2007), more highly developed phonological awareness after sustained bilingual education (Laurent & Martinot, 2010), and more rapid growth in vocabulary knowledge (Lo & Murphy, 2010). A study in South Africa (De Sousa 2012) compared reading and understanding between Afrikaans-English balanced bilinguals and Zulu-English partial bilinguals who went to bilingual schools, with English monolinguals who went to monolingual English schools. The Afrikaans-English bilinguals showed similar performance on English decoding as English monolinguals, but better performance on comprehension tests in English. Afrikaans-English bilinguals outperformed Zulu-English bilinguals on both measures. As with much of this type of research, there are other explanations as to why children from communities where Zulu is spoken might not perform as well as children who come from communities where Afrikaans is spoken. Aldosari and Alsultan (2017) compared home language literacy development of children in bilingual schools in Saudi Arabia and similar children in monolingual Arabic schools. They found that there was no detriment to the children’s Arabic literacy associated with attending the bilingual schools. In a study by Taylor and Lafayette (2010), 3rd to 5th Grade children attending bilingual schools in Louisiana, USA, were compared to 3rd to 5th Grade children from the same school districts attending monolingual schools. They assessed these children on a number of different test batteries. Children at the bilingual schools outperformed their monolingual peers in tests of English language arts and language achievement tests. The other findings from this study will be discussed elsewhere in this document. Finally, in the Fox et al. corpus on bilingual schooling and literacy is a study of dual language Mandarin-English immersion (Padilla, Fan, Xu & Silva 2013). Forty kindergarteners (20 English first language and 20 Mandarin first language) were followed until 5th Grade, through a small dual language Mandarin-English programme which existed within an otherwise monolingual English school. The participants’ scores on state mandated English language arts tests at Grades 2 and 3 were lower than for children in the monolingual part of the school. However, by Grade 5 they had overtaken their monolingual peers and were statistically significantly out-performing them on these tests. The authors emphasise that dual language instruction sustained over many years was the key to the children's success.

We found two additional studies that met Fox et al.'s inclusion criteria for studies of the relationship between foreign language learning and literacy. Aljohani (2016) compared Saudi Arabian children's scores on tests of Arabic writing, grammar and reading between children attending private schools, where English is taught, and those attending government schools, where English is not taught. The author says that there were big differences between scores from these two types of school but offers no statistical analysis to back up these claims. In addition, serious potential confounds exist relating to school type (private compared to government and all girls compared to all boys). Festman and Schweitzer (2019) tested bilingual and monolingual children on reading comprehension, reading fluency and spelling. They found no difference between reading comprehension and reading fluency scores between groups. The monolinguals outperformed the bilinguals on spelling tests.

3.4.7.2 Other academic attainment

In four studies in the Fox et al. review the effects of knowing or studying a foreign language on maths attainment was explored. Garrett (2011) analysed state level data on mathematics achievement for children in states in the USA where bilingual education had, historically, been mandated for language minority children (Arizona, California, Massachusetts and Texas), but which was withdrawn around the turn of the millennium in favour of English only education. Analysis of the patterns of attainment before and after this policy change suggests either no impact of the change or a reduction in maths attainment for students who would otherwise have been placed in bilingual programmes. In a study of adults, Onnis, Chun, and Lou-Magnuson (2018) found that bilinguals had better core statistical understanding than monolinguals. Zaunbauer and Möller (2010) compared maths attainment of German Grade 1 and 2 children in a bilingual school (where all subjects, including maths, were taught in English, except reading and writing which was taught in German) with similar children at a German monolingual school. Maths attainment was similar in both groups, but children in the bilingual programme made more rapid progress between Grade 1 and Grade 2. It seems likely, however, that this rapidity is reflective of a slower start while children were developing competence in English, followed by an uptick once their proficiency in English had improved.

In a related study (Kuska, Zaunbauer & Möller 2010), German children in bilingual and monolingual programmes were taught using standardized approaches. Those attending the bilingual programmes showed better learning and memory performance than their peers in monolingual schools. Bialystock and Feng (2009) found a similar effect of bilingualism on memory. Among the test batteries used by Taylor and Lafayette (2010) to compare children in Louisiana State bilingual programmes with similar children at monolingual schools was the Louisiana Educational Assessment Program for the 21st Century (LEAP 21) test. Children at bilingual schools outperformed those at monolingual schools to a highly statistically significant degree on measures of mathematics, science, and social studies. Finally from the Fox et al. literature was a study of ninth grade Tatar children in Russia. The study used regression analysis to explore the relationships between course grades and bilingualism. Tartar children who spoke Tartar at home (and who were, therefore, classified by the authors as bilingual) were more likely to excel at chemistry and literature than Tartar children who spoke Russian at home (and who were therefore classified as not-bilingual by the authors).

3.4.7.3 Impact on standardized test scores

A small body of literature on the effects of foreign language learning/bilingualism on standardized test scores was located by Fox et al. for their review. Cooper et al. (2008) analysed the SAT Reasoning scores of more than 9,000 high school students in Atlanta, Georgia, USA, comparing the results of students who had taken a foreign language course and those who had not. They found that students who had taken a foreign language course statistically significantly outperformed those who had not on the SAT. This difference was more pronounced for students who had taken the course for longer periods. However, because taking the course for longer periods relied on successfully completing the courses at lower grades, there is the possibility that survival bias played a part in these results.

In the study by Taylor and Lafayette (2010), already discussed, part of the test battery included the IOWA Test of basic skills. As with the other findings in this study, children who had learned another language in the early years, through either being born into bilingual households or attending bilingual schools, outperformed monolinguals on all measures.

We located one additional study that met the inclusion criteria for Fox et al.'s review. Gunzenhauser, Karbach and Saalbach (2019) compared the verbal competencies of 21 3rd Grade bilingual children with 26 monolingual children, in Germany. Monolingual children did moderately but statistically significantly better than bilingual children on standardized tests of verbal competence. In one other there were no differences between monolingual and bilingual children on other tests used in the study.

3.4.8 Creativity

There is a suggestion in the literature that bilingualism and/or learning a foreign language can have positive effects of creativity. Seven primary studies were identified by Fox et al. addressing this theme. Kharkhurin (2009) assessed monolingual Farsi speakers living in Iran and bilingual Farsi-English speakers living in the UEA on tests of divergent thinking and structured imagination and found that bilinguals scored higher than monolinguals. The author cautions that there may be uncontrolled-for differences between groups that may explain the differences. In a later study, Kharkhurin (2010) assessed the performance of Russian-English bilingual and English monolingual college students in the USA on the verbal and nonverbal indicators of the Abbreviated Torrance Test for Adults. The results showed an advantage for bilingualism in non-verbal creativity and an advantage for monolinguals in verbal creativity.

Lee and Kim (2010) explored relationships between the degree of bilingualism and creative styles and strengths among 7-18 year-old Korean-American students. Participants were assessed for degree of bilingualism and grouped into monolinguals, unbalanced bilinguals and balanced bilinguals. The authors found no statistically significant differences in mean scores on tests of innovative and adaptive creativity between groups. A study by Hangeun and Hee Kim (2011), which argued for positive relationships between degree of bilingualism and creativity, was reported in Fox et al.'s review. However, this study has been exposed as reusing data from the above study by Lee and Kim (2010), manipulating those data to redraw the boundaries between monolingual, unbalanced bilingual and balanced bilinguals in different places, without making that clear in the publication. Offered the opportunity to provide a corrigendum to the 2011 paper to make clear the re-use of old data and how these data had been manipulated differently, Hangeun and Hee Kim refused. We therefore reject the findings of this study on methodological grounds and poor scientific practice. The degree of bilingualism and its relationship was also assessed by Kostandyan and Ledovaya (2013). They found that simultaneous bilinguals statistically significantly out-performed sequential bilinguals in tests of nonverbal flexibility.

Divergent thinking was assessed in monolinguals and advanced EFL learners in studies by Fürst and Grin (2018) and Ghonsooly and Showqui (2012). Findings from Fürst and Grin's correlational analysis of 596 adults suggest that foreign language learning is positively related to divergent thinking. In comparing advanced teenage EFL learners with beginner EFL learners aged 16-18 on scores on Torrance Test of Creative Thinking, Ghonsooly and Showqui found that advanced learners outperformed beginners statistically significantly on all four measures of divergent thinking assessed.

3.4.9 Communicative and intercultural competence

Eight studies that addressed intercultural and communicative competence were located and synthesised by Fox et al. One study (Collins, Toppelberg, Suárez-Orozco, O'Connor & Nieto-Castañon, 2011) found that the ability to speak another language was associated with emotional, social, and behavioural well-being among children of immigrants (which suggests perhaps that this may not have been intercultural competence per se, rather, that being able to speak the language of the country to which one migrates is helpful). In a similarly uncontroversial finding, Domínguez and Pessoa (2005) assert that early learning of a foreign language supports oral skills and confidence in using that language. Dewaele (2010) suggests that an intermediate facility with a foreign language (but not weak or strong proficiency) is associated with better navigation of challenging communicative interactions. These data are based on self-report of participants' communicative competence.

Mikolic (2010) compared communicative competence among Slovene individuals across the population age. Younger individuals were found to be more communicatively competent than older individuals. This finding is attributed to the increase in opportunities for bilingual Italian-Slovene education and exposure to Italian media since Italian speakers were officially granted minority status in the region.

A study of intercultural competence (Hoyt 2016) among university students attending French conversation classes in the USA found, on the basis of questionnaires, that their skills of interpreting and relating and their critical cultural awareness improved over time. Two more uncontroversial findings are reported by Jiang and Wang (2018) and Barski and Wilkerson-Barker (2019). In the first, questionnaire data was used to assert a strong positive relationship between years spent learning a foreign language and communicative competence in that language. The extent of communicative competence was also positively correlated with measures of intercultural empathy. In the second, university students in a one-semester beginner foreign language course demonstrated no significant change in intercultural competence. The authors suggest that this was because they were concentrating on learning the language, and that had explicit instruction in intercultural competence been included in the course, results might have been different.

In the only experimental study in this literature, Coelho, Andrade and Portugal (2018), 16 Portuguese pre-schoolers were assigned to take part in a three-month programme called Awakening to Languages (AtL) or to continue with business as usual. The AtL programme had a positive impact on oral comprehension and attitudes towards language and cultural diversity.

We found two additional studies that addressed this theme. Fecher and Johnson (2018) showed babies from bilingual households and monolingual households two videos of people speaking Spanish (a language unfamiliar to them). Later they showed the same videos, but the voices of the speakers had been switched. That is, the voice of Speaker 1 was paired with the image of Speaker 2 and vice versa. Babies from bilingual households noticed the switch, whereas monolingual babies did not (evidenced by longer gaze times by the bilingual babies). The authors suggest that Bilingual infants are more sensitive to talker identity information encoded in the acoustic speech signal than monolingual infants.

Lorge and Katsos (2018) had bilingual and monolingual adults explain how to make an apple, peach and blueberry cake to a monolingual English child, a monolingual English adult, and an adult whose first language was Greek and whose English was heavily accented (though otherwise of very high English proficiency). They found that bilinguals made more attempts to facilitate communication when speaking to the child and the heavily accented adult (they employed a wider range of pitch when speaking to the child and hyper-articulated vowel sounds when speaking to the Greek adult) compared to monolinguals. However, measures of a total of ten factors were assessed but only two were reported in full. In addition, a number of ANOVAs were conducted on the data with no adjustment for multiple tests. Both of these facts suggest that data dredging and Type 1 errors cannot be ruled out.

3.4.10 Aging and health

An area of interest in the role of additional language learning and bilingualism that has made headlines recently is its potential effects on aging and health, in particular the effect of learning or knowing another language on the onset of dementia and Alzheimer's. Fox et al. found 13 studies that addressed this phenomenon. The majority of these studies found that language learning and bilingualism are associated with positive outcomes in these domains. Importantly, however, none of these studies looked at the effects of teaching a foreign language, rather they examined differences between already bilingual adults and monolingual adults. In some of these bilingual participants their bilingualism was sequential (that is, the second language had been learned in school after they had already acquired their first language). In others, they were simultaneous bilinguals (that is, they were exposed to two or more languages since birth, or very early childhood). This makes interpreting the data difficult as there could be competing explanations for the differences in outcomes that are unrelated, or only tangentially related, to bilingualism. That said, Fox et al. found evidence that individuals with Alzheimer's who had been lifelong bilinguals or who had learned another language in their early schooling and had continued to use it showed delays in memory loss and delayed onset of Alzheimer's compared to monolinguals (Chertkow et al. 2010; Craik, Bialystok, & Freedman 2010). Similarly, in studies of older adults Del Maschio et al., 2018; Perani et al., 2017; Woumans, Santens et al., 2015) all reported delays in the onset of dementia and cognitive decline in bilingual adults relative to monolinguals. Other studies found evidence of increased selective attention skills (Salvatierra & Rosselli 2011), and higher cognitive reserves in advanced age (Bak, Nissan, Allerhand, & Deary, 2014; Ihle, Oris, Fagot, & Kliegel, 2016; Perquin et al., 2015; Woumans, Santens et al., 2015).

Not all studies that were included in Fox et al.'s review were positive. Zahodne, Schofield, Farrell, Stern, and Manly (2014) compared dementia diagnosis rates among bilingual and monolingual adults in the USA and did not find a difference. Kousaie and Phillips (2012) compared non-immigrant bilinguals in the USA with monolingual counterparts on Stroop tests (tests of attentional control) and did not find an advantage for either group. Liu, Liu, Yip, Meguro, and Meguro (2017) compared rates of cognitive decline and dementia in bilingual and trilingual Taiwanese adults in the assumption that more languages might be associated with better effects. They did not find a difference in outcomes between groups.

We located two additional studies that met Fox et al.'s inclusion criteria for the effects of bilingualism on health and aging. Paplika et a. (2019) examined characteristics of bilingual and monolingual adults in India suffering from aphasia following ischemic stroke. Sixty-five patients with stroke related aphasia (27 monolinguals and 38 bilinguals) were assessed for stroke characteristics and aphasia severity. They found that differences between groups in stroke characteristics, age, occupational status, frequency of vascular risk factors, presence of previous strokes, location and laterality of infarcts, and time after stroke were statistically non-significant. Analysis of the severity of aphasia, however, revealed that bilinguals scored statistically significantly higher on measures of language and fluency, attention, memory, and visuo-spatial skills. The authors suggest that the bilingual group's bilingualism offered a protective effect on the severity of their aphasia, if not against the frequency of the condition. It is important to note that the level of education among the bilingual group was statistically significantly higher than that of the monolingual group.

Papageorgiou et al. (2019) set out to assess the claim that lifelong bilingualism has protective effects against cognitive symptoms of aging. Thirty-seven bilingual and 37 monolingual healthy British adults aged about 70 years old were recruited to the study and matched on age, gender and socioeconomic status. Bilingual participants were highly proficient in at least two languages. They all reported using both languages on a daily basis for more than 50 years. All monolingual participants reported little or no exposure to any language other than English. They were submitted to a battery of six tasks: English vocabulary knowledge (British Picture Vocabulary Scale III [BPVS III]), non-verbal reasoning (Raven's Advanced Progressive Matrices), executive function (Simon Task), planning and problem solving (Tower of London), verbal working memory (Digit Span forwards and backwards), and visuo-spatial working memory (Change Blindness). Both groups performed comparatively on all tasks, with the exception of the Tower of London task, where monolingual participants were faster than bilinguals in deciding the first move.

3.4.11 Positive attitude toward other languages/cultures

Three studies included in the Fox et al. review reported positive effects of early foreign language learning and bilingual education on attitudes towards other languages and other cultures (Heining-Boynton & Haitema, 2007; Hood, 2006; Merisuo-Storm, 2007). Heining-Boynton & Haitema (2007) used questionnaires to probe attitudes towards Foreign Language in the Elementary Schools (FLES) programmes (in which Spanish and French were taught using mainstream curriculum content as the vehicle for language learning) in North Carolina, USA, among children from a variety of linguistic and cultural backgrounds. They then followed this up ten years later with structured interviews with the participants. At both time points data revealed positive attitudes towards language learning, foreign language speakers and foreign cultures. The study by Merisuo-Storm (2007) among Finnish 4th Graders revealed more positive attitudes toward foreign language learning for children enrolled in bilingual classes (where 20% of instructional time was conducted in English) compared to similar children in monolingual Finnish classes. The study by Hood (2006) asserted that commitment to early foreign language learning in the school as a whole raised attitudes towards other cultures.

We located one additional study (Singh et al, (2019)) which used a quasi-experimental design to explore the reactions of infants from bilingual and monolingual households to speakers of the same or different race as them. Stratified by bilingualism or monolingualism, infants were randomly allocated to one of four conditions. In all conditions they watched a video of an adult encouraging them to look for a puppet in a particular region of the screen. In one condition the adult was reliable, that is, the puppet was where the adult said it was. In the other condition the adult was partially reliable, that is the puppet was sometimes where the adult said it was and sometimes it was not. The conditions also differed by the race[8] of the adult. For some infants the adult was the same race as them, for others the adult was of a different race. On the basis of gaze tracking, the study found that bilingual children did not

[8] Race is the term used in the original study.

discriminate their levels of trust in either wholly or partially reliable conditions on the basis of race. Monolingual children, however, demonstrated equal levels of trust in wholly reliable adults regardless of race, but discriminated by race in the partially reliable conditions. That is, they did not trust the different race adult as much as the same race adult.

3.4.12 Employability and earning potential

Often celebrated as a reason for learning or knowing another language is the assumption that it raises the employability of the individual. Facility with another language is colloquially seen as an asset that enhances job prospects. Fox et al. found ten studies that addressed issues of employability. Only two of these reported objective data on wage differentials (Saiz and Zoido 2005; Godoy et al. 2009). Using regression analyses, Saiz and Zoido (2005) found that college graduates in the USA who could speak another language had slightly higher wages than those who could not. Godoy et al (2009) looked at employment rates of Bolivian minority speakers of a regional linguistic variety called Tsimané. Tsimané speakers who were also highly proficient in Spanish, the language of the majority, earned substantially more than monolingual Tsimané speakers and Tsimané speakers with only moderate proficiency in Spanish. The remaining studies located by Fox et al. are all either qualitative in nature (ethnographies and case studies) or collected only perceptions and opinions via questionnaires. They all converged on the overall finding that employers and employees view the ability to use another language as an advantage (Millar 2017; Makumane & Ngcobo 2018; Belpoliti & Pérez 2019; Claassen, Jama, Manga, Lewis, & Hellenberg 2017; Beadle, Humburg, Smith, & Vale 2016; Duran 2016; Gogonas & Kirsch 2018). These studies are not able to say whether this observed enthusiasm for other languages translates into improved job prospects. Potentially instructive on this point, however, was a study conducted by Damari et al. (2017). They surveyed over 2,100 employers in the United States on their employment practices around multilingual employees. They found that 93% said that they valued employees who could work effectively with clients, customers and businesses from a range of countries and cultures. However, only 10% said that they required new employees to be able to use another language, only 41% said that facility with another language were an advantage when hiring new people, and only 66% asked about additional language proficiency during the hiring process. This suggests that while employers might say they value the ability to use another language among their employees, it appears that this may not be a crucial factor in their hiring decisions.

3.4.13 Motivation

Two studies in the Fox et al. review explored the relationships between learning another language and motivation (Brumen 2011; Hood 2006). These were both observational studies. The first found that Slovenian school children aged between 4 and 6 learning German and English associated high levels of motivation with learning these languages, citing enjoyable communicative oriented lessons and a positive, supportive classroom atmosphere as reasons for this. The authors argue that this motivation to learn extended beyond the MFL classroom into motivation to learn in other curriculum areas. The second study noted that a positive attitude toward languages embodied in the ethos of the school reflected a generally well-motivated student body.

We did not locate any additional studies that focussed on motivation and language learning.

3.4.14 Summary and Conclusion

There are many dimensions addressed by the Fox et al (1 and 2) (2019) reviews ranging from the impact of bilingualism on the architecture of the brain to whether knowing and/or using another language leads to greater employment possibilities. We summarise the key findings emerging from Fox et al's (2019) reviews and our updates below in Table 3.5.

Table 3.5. Summary and conclusions following from the Fox et al (1 and 2) updates

Topic	Implication	Additional Comment	Studies Contributing to this conclusion
Cognitive abilities	There is some mixed evidence to suggest advantages for bilinguals on executive functions, inhibitory control, working memory, attentional control and cognitive flexibility.	The research here is variable – for every study purporting bilingual advantages there seem to be some suggesting otherwise. There is also variability in quality of research. This means it is difficult to reach a conclusion here. We would argue that benefits of bilingualism (being able to speak more than one	Bialystok (2007); Abdelgafar & Moawad (2015); Adi-Japha et al (2010); Antón, et al (2019); Babcock & Vallesi (2017); Blom et al (2017); Blom, et al (2014); Blumenfeld & Marian (2011); Carlson & Meltzoff (2008); Choi et al

		language) is tangible and of greatest interest from educational perspectives.	(2018); Chung-Fat-Yim, et al (2017); Cockcroft et al (2019); Colzato et al. (2008); Comishen et al. (2019); Costa et al (2009); Damian et al (2019); Fernández et al (2013); Festman et al (2010); Hernandez et al (2010); Jiao et a; (2019); Kalia et a; (2019); Karlsson et al. (2015); Kazemeini & Fadardi (2016); Kousaie et al (2014); Kuipers & Thierry (2013); Kushalnagar et al (2010); Ladas et al (2015); Macnamara & Conway (2014); Marini et al (2019); Marzecová et al. (2013); Ong et al (2017); Paap et al (2019); Paap et al (2014); Papageorgiou et al. (2019); Park et al (2018); Poulin-Dubois et al (2011); Prior & MacWhinney (2010); Robinson & Sorace (2019); Ross & Melinger (2017); Schroeder & Marian (2012); Soveri et al (2011); Stafford (2011); Sun et al (2019); Tran et al (2019); Valis et al (2019); Vega & Fernandez (2011); von Bastian et al (2016); Wodniecka et al (2010); Woumans et al (2019); Yang & Yang (2016).
Linguistic processing and reasoning	Some evidence bilinguals have enhanced	As above, the evidence is too variable in terms of	Choi, Rouse & Ryu (2018); Jaekel et al

	conversational competence, but again, mixed evidence regarding advantages in vocabulary, reading and pragmatics	outcomes and methodological rigour to be able to reach a definitive conclusion about a bilingual advantage on these dimensions	(2017); Kaushanskaya et al (2011); Marinova-Todd (2012); Siegal et al. (2010); Thompson (2013); Verhagen et al (2017).
Metalinguistic Awareness	Evidence to indicate advantages for bilinguals on phonological, morphological and syntactic awareness	The evidence here is more consistent and robust indicating bilingualism can enhance individuals' ability to see aspects of language as entities within their own right. This is important as metalinguistic awareness is associated with literacy skill.	Dillon (2009); Ibrahim et al (2007); Kuo & Anderson (2010); Schoenpflug & Klische (2010); Verhoeven (2007); Bialystok & Barac (2012); Diaz & Farrar (2018); Bien-Miller et al (2017); Daller & Ongun (2018); Hermanto et al (2012); Reder et al (2013); Sun et al (2018); Yeon et al (2017); Laurent & Martinot (2010); Huang (2018); Barbier & Niit (2006).
Cognitive Development	Again, mixed evidence here suggesting the cognitive architecture may or may not be different for bilinguals	Given more recent research is questioning bilingual advantages in cognitive development, it is difficult to reach any conclusions here as of yet.	Bialystok & Craik (2010); Lazaruk (2007); Arredondo et al (2017); Bartolotti et al (2017); Jasinska & Petitto (2013); Kuhl et al. (2016); Santillán & Khurana's (2018); Folke et al (2016); Struys et al (2018); van Veen et al. (2019).
Cross-Language Activation	There is evidence for enhanced language switching in bilinguals	Given bilinguals have, by definition, two (or more) languages, it is no surprise they are better at task switching within languages and it is not clear what educational advantages this might confer.	Costa et al (2006); Kormi-Nouri et al. (2008).
Spatial reasoning	There is some evidence that bilinguals have enhanced spatial reasoning	At present there is less evidence available here and consequently further work is needed.	Greenberg et al (2013); Stephens & Moxham (2019).

Academic achievement	There is evidence that bilinguals have some advantages in aspects of literacy, mixed evidence as to whether bilinguals have advantages in maths, and some evidence students who have taken a FL have higher standardised test scores (SAT).	There is insufficient evidence in this domain to make firm conclusions. The most promising is arguably the literacy domain due to the more robust findings that bilinguals have advantages in metalinguistic awareness, but more work is needed in these, and other, academic domains	Aldosari & Alsultan (2017); Aljohani (2016); Bialystok & Feng (2009); Cooper et al. (2008); Festman & Schweitzer (2019); Garrett (2011); Gunzenhauser et al (2019); Knell et al. (2007); Kuska et al (2010); Laurent & Martinot (2010); Lazaruk (2007); Lo & Murphy (2010); De Sousa (2012); Modirkhamene (2006); Onnis et al (2018); Padilla et al (2013); Silven & Rubinov (2010); Taylor & Lafayette (2010); Zaunbauer & Möller (2010).
Creativity, intercultural competence	These areas have mixed evidence to support claims for bilingual advantages here. The most consistent evidence is in the area of communicative and intercultural competence	There is some evidence that bilinguals have higher scores on divergent thinking tasks and on measures of intercultural awareness but as with other areas in this research, more work is needed.	Barski & Wilkerson-Barker (2019); Coelho et al (2018); Collins et al (2011); Dewaele (2010); Domínguez & Pessoa (2005); Fecher & Johnson (2018); Fürst & Grin (2018); Ghonsooly & Showqui (2012); Hangeun & Hee Kim (2011); Hoyt (2016); Jiang & Wang (2018); Kharkhurin (2009); Kharkhurin (2010); Kostandyan & Ledovaya (2013); Lee & Kim (2010); Lorge & Katsos (2018); Mikolic (2010).
Ageing and health	Some research has suggested that being bilingual can delay the onset of dementia/Alzheimer's and general cognitive decline, but other research finds no such conclusion	As with other domains, this research area does not have sufficient, consistent evidence to reach a firm conclusion either way.	Bak et al (2014); Chertkow et al. (2010); Craik et al (2010); Del Maschio et al. (2018); Ihle et al (2016); Kousaie & Phillips (2012); Liu et al (2017);

			Papageorgiou et al. (2019); Paplika et a. (2019); Perani et al. (2017); Perquin et al. (2015); Salvatierra & Rosselli (2011); Woumans et al. (2015); Zahodne et al (2014).
Employability, motivation	A small set of research suggesting bilinguals are more employable, and that knowing another language is associated positively with motivation	Again, there is simply not enough evidence examining these relationships to be able to reach firm, definitive conclusions about the relationship between bilingualism and earning potential or motivation	Beadle et al (2016); Belpoliti & Pérez (2019); Brumen (2011); Claassen et al (2017); Damari et al. (2017); Duran (2016); Godoy et al. (2009); Gogonas & Kirsch (2018); Hood (2006); Makumane & Ngcobo (2018); Millar (2017); Saiz & Zoido (2005).
Attitudes towards other cultures	A small number of studies suggest that knowing or learning another language is associated with positive attitudes towards other cultures.	While all studies converge on similar findings, the size of the body of literature, methodological quality, and the potential influence of confounders associated with learning other languages mean this is far from a definitive finding.	Heining-Boynton & Haitema (2007); Hood (2006); Merisuo-Storm (2007); Singh et al. (2019).

Taking the findings of Fox et al.'s reviews together with the additional studies we identified, a mixed picture of the wider implications of bilingualism/learning another language emerges. There may be some advantages to having learned another language beyond the self-evident ones. These include meta-linguistic awareness, intercultural competence, and some cognitive functions. However, the trustworthiness of these findings is questionable on the basis of methodological shortcomings in the literature, selective reporting, and the inherent difficulty in disentangling bilingualism *per se* from the social contexts in which it arises. It is therefore difficult to draw firm conclusions either way on the basis of the evidence identified.

We began this chapter by referencing Murphy et al (2015) which clearly demonstrated that learning a FL can have advantages on developing L1 literacy and the evidence from the reviews supports this finding where metalinguistic awareness is very clearly supported by bilingualism. If there is genuine interest in whether learning another language improves other aspects of a learner’s education, then fair tests comparing the effects of foreign language learning on other academic outcomes must be conducted in greater number. Fox et al. echo this sentiment in their conclusion, noting that much of the research on this question comes from the fields of cognitive sciences, psychology, and neuropsychology rather than education. Fruitful research partnerships between these fields are thus ripe for the picking.

3.5 Review Questions 4 and 5

RQ4: What is the impact of using a non-native language as the medium of instruction in academic subjects on students' academic outcomes?
RQ5: Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?

In this chapter we present the findings of our investigation into the fourth and fifth review questions. We have taken these two questions together in this chapter as RQ5 is directly related to the focus of RQ4. We first describe the seed reviews used to address these two RQs and then present our updates to these reviews. We present our conclusions by summarising the evidence and offering our critical evaluation at the end of this section.

On examining the four systematic reviews which were included in our initial trawl in answer to these two review questions (Graham et al 2018; Fitzpatrick et al 2018; Lo & Lo 2014; Goris et al 2019) we encountered problems of definitions surrounding approaches to medium of instruction as well as diversity of systematic review objectives. First, the four reviews referred differently to the classrooms they were investigating. Graham et al (2018) use an 'an umbrella term' of 'Content-Based Instruction' (CBI) even though their included studies used either 'Content and Language Integrated Learning' (CLIL), or 'English Medium Instruction' (EMI). Fitzpatrick et al. (2018) refer to them as CLIL classrooms as do Goris et al. (2019). Lo & Lo (2014) label these classes as 'English Medium Instruction' (or English Medium Education).

This differential terminology may lead to a problem in assembling an overall picture as it stems from the geographical perspectives in which the authors were themselves working in: CBI tending to be used in North America (even though the studies that Graham et al. reviewed were not situated in North America); CLIL in Europe; and EMI in Asia. Different geographical locations can also reflect different practices: The educational settings of the European studies examined might vary in terms of the number of subjects studied in the L2, and even whether one subject was entirely being taught in the L2 or for only a few hours per week. In Hong Kong, for example, (the entire setting of the Lo & Lo review) most of the subjects in some secondary schools are taught through the L2. We should also note that—as far as we can glean from the reports—all of the studies examined were using English as the L2 medium of instruction. In other words, collectively these reviews cover academic subjects being (partly or fully) taught through the medium of the current international language of communication. This has implications for whether the findings of these reviews can be transferred to CLIL classrooms in Anglophone countries such as the UK where French, Spanish, German or Mandarin might be the L2 medium of instruction.

There were also different objectives in the systematic reviews. Graham et al. (2018) were interested in whether using the L2 had an impact both on language and content, as did Lo & Lo (2014). Goris et al. (2019) and Fitzpatrick et al. (2018) only observed the impact of using the L2 on English language learning.

Lastly, we should note that some of the reviews refer to CLIL as 'a method' of teaching (see Fitzpatrick et al. 2018) whereas EMI is rarely referred to as a 'method' but more as a policy decision enacted by institutions or governments. Furthermore, to what extent CLIL can be described as a teaching method, which requires pre-defined procedures, is open to question. The descriptions provided in individual CLIL studies rarely refer to a systematic application of guidelines as to how to teach both content and language and elsewhere we can point to evidence that at least some teachers wonder whether they are teaching CLIL at all (Di Martino & Di Sabato 2012).

3.5.1 Synthesis of seed reviews

We now summarise the findings of these four reviews in order to then be able to update the findings with the results of our own systematic search.

In broad terms Graham et al. (2018) investigated 'how CBI teaching practices compare with traditional language teaching' (p. 21). Caution should be exercised here as 'traditional language teaching' with which CBI is being compared might vary hugely from high L2 input communicative approaches to low L2 input grammar-translation, casting doubt on the validity of the outcome measures being adopted. Nonetheless, Graham et al. (2018) base their review on two theoretical underpinnings: the input hypothesis (that the L2 can be understood and even learnt through the L2, provided that the learner is receiving input at no more than the level just above their current linguistic knowledge) and cognitive load theory (that learners' focus may be diverted either, in our case, towards understanding and learning content or comprehending the language in which that content is being delivered). Put differently, the review explores whether CLIL/EMI students learn from high exposure to L2, and investigates the possibility of learning being compromised by students needing to simultaneously focus on both subject content and on the language being used to deliver that content.

Graham et al. (2018) report on 25 CBI studies which were included after criteria were applied (22 originally labelled CLIL and three EMI) of which 19 examined language outcomes and 6 content outcomes. They found that some studies reported an advantage for CBI over EFL in both language skills and in content learning, while in others there was no advantage. In the case of content learning, where the comparison is between content being learnt in the L2 versus the L1 (see Lo & Lo 2014), there were some disadvantages for CBI especially with 'low achieving students' (p. 32). However, CBI showed no significant disadvantage against students taught via EFL in terms of language learning.

The review authors' overall conclusion is that the results are mixed and that even those that found an advantage for CBI have a number of methodological problems associated with them: in many CBI programmes students (or their parents) had elected to enrol on this type of programme rather than an EFL programme. The self-selection of students on these programmes raises the possibility that they had greater motivation, or greater initial language proficiency or both. In some studies, CBI students also had extra instruction time. These imbalances might favour CBI outcomes over EFL outcomes. Conversely the review's finding that the outcome measures used in the studies were evaluations of general everyday English language proficiency, might favour EFL classes over CBI classes as in the latter there is likely to have been much greater focus on academic language. Although the authors mention the methodological problem of the "wide variety of CBI instruction in different schools" (p. 31) being compared, they also put emphasis on the possibility that variety similarly exists across EFL instructional contexts, and thus these contexts are broadly comparable.

In the systematic review carried out by Fitzpatrick et al. (2018), the focus was entirely on whether CLIL classrooms outperformed language-only classrooms in terms of language learning. The vast majority of these studies were carried out in Spain providing advantages in terms of uniformity of context but disadvantage in terms of generalizability across geographical settings. Moreover, we should note that one of the aims of their review was to inform the Welsh language context in which a debate exists as to whether some content should be learnt in Welsh, perhaps making the Spanish comparison (with its multilingual nature) more appropriate.

The authors reviewed six studies which focused on vocabulary learning. They found a tendency for CLIL learners to perform better than EFL learners. The three studies reviewed which focused on grammar, found mixed results or relatively small advantage for CLIL, whilst some advantage was found for CLIL classrooms in terms of reading comprehension and for listening and speaking. Nevertheless, Fitzpatrick et al (2018) comment that "it is difficult to draw any strong conclusions in terms of the direct effect of CLIL as in most studies the amount of exposure to language was greater than for the control non-CLIL groups " (p. 60) . Similar to Graham et al (2018), throughout their review they repeatedly emphasise that any CLIL advantage cannot be attributed to the CLIL approach alone, due to these group differences.

Like Graham et al (2018), the meta-analysis by Lo and Lo (2014) investigated the effects of EMI compared with Chinese Medium Instruction (CMI) on both language and content. The educational setting for this review was restricted to Hong Kong. However, because of the historical impact of British colonialism, the review was able to go back as far as 1970, leading to a final sample of 24 studies included meta-analysis. Furthermore, they included EMI's impact on L1 Chinese, which was an outcome variable not considered in other reviews.

In terms of content learning (academic achievement) in subjects such as maths, science history and geography, this meta-analysis found that EMI students do not compare favourably with CMI students although the deficit was not significant in maths. They also 'lagged behind' (p. 57) their CMI peers in terms of their L1 development although the combined effect size in L1 development was not high. However, EMI students, in the studies they reviewed outperformed the CMI students in English language achievement with moderate effect size and also demonstrated a higher self-concept. They conclude that “findings imply the fact that EMI students in Hong Kong may have sacrificed academic achievement for L2 proficiency” (p. 63).

However, Lo and Lo (2014) point to a number of moderator variables which need to be considered when evaluating these findings, particularly the outcome measures used. Here it was found that EMI students slightly outperformed CMI students on content achievement when standardised tests were used, but comparatively underperformed when ‘self-designed tests’ were used. The authors attribute this to the latter being more specifically designed to assess learning on certain topics taught during the research period” (p. 63). It could therefore be possible that L1 medium of instruction benefits short term content learning and that EMI learners might catch up in the long term (see Marsh et al., 2002).

Another moderator variable they highlight is the point at which EMI education begins. The researchers question whether switching from L1 medium of instruction to EMI at grade 7 is too early in terms of students’ L2 proficiency levels at the end of elementary school. Lastly, they emphasise the moderating effect of the mother-tongue policy implemented (somewhat unsuccessfully) in the late 1990s, which resulted in EMI students becoming much fewer in number, leading them to outperform CMI students because of the higher levels of selection involved in being admitted to EMI schools. In other words, once again, the issue of prior language proficiency, selection and enrolment need to be considered carefully when attempting a broad assessment of the effectiveness of EMI and CLIL.

Similar concerns about methodology are expressed from the outset by Goris et al (2019) who argue that although many studies have shown benefits for CLIL, these are frequently attributed by commentators to the attraction to CLIL or the selection of high-achieving or highly motivated students as well as to longer periods of exposure. These authors provide overviews of 4 primary school studies and 18 secondary school studies, of which many investigate vocabulary. In the primary sector, the findings show an advantage for vocabulary with quite high effect sizes (although these do not appear to be pooled by the review authors). In the secondary sector the findings of studies are more mixed.

They summarise their findings by pointing to the fact that, apart from vocabulary learning, such a variety of competences have been investigated as to be unable to draw concrete conclusions. As a result, there is a lack of depth of research in the field that can provide reliable answers to the review question. They conclude that there is no unequivocal evidence that a CLIL approach leads to significant gains in English language proficiency. However, we wish to pick up on one part of their discussion which we feel needs further development. Goris et al. (2019) found three studies which reported significant effects for CLIL in terms of spoken fluency. They hypothesise that this was due to "the increased opportunity for authentic communication" (p. 693). We believe this hypothesis needs to be challenged on both theoretical and empirical grounds. At a theoretical level, there is no reason to suppose that content classrooms (whether they be in CLIL, EMI or L1) provide any more opportunities for learners to speak than EFL classrooms. In content-driven education there may be a great deal more 'subject' matter (e.g. in science) for the teacher to put across than in language-only classrooms. This is supported by empirical evidence in many studies of EMI/CLIL classroom interaction (e.g. Lo 2012 and Pun & Macaro 2019 in Hong Kong; Macaro et al. 2019 in Italy) which do not show high levels of student oral participation. This brings us back to the question we raised earlier: to what extent is it possible to simply compare language outcomes of EMI/CLIL classrooms with language-only classrooms without controlling (in addition to the selection and extra EFL issues raised above) for the kind of pedagogy which is being adopted in both types of classrooms?

3.5.2 Updates to the review

In order to update the four seed reviews described in brief above, we carried out a systematic review of articles published since each of the seed-reviews collected their data. This allowed us to update the findings of these reviews and further answer the question of whether using a non-native language as the medium of instruction in academic subjects impacts on students' academic outcomes. The seed review of Goris et al. (2019) revealed no new studies that fit the inclusion and exclusion criteria of the original reviews and met the requirements on our rapid assessment evaluation. The seed reviews of Fitzpatrick et al. (2018) revealed two new important studies (rated 3* or above); and Graham et al. (2018) revealed eight new studies, of which three were evaluated of high importance (3*) to our research question. Lo and Lo (2014) revealed four additional studies, of which two were evaluated to be of greater importance in contribution (3*). In the update of Harris & Ó Duibhir (2011) two studies were identified which met the inclusion criteria of this research question. These were Berens et al (2013) and Lo and Murphy (2010). However, these were ultimately discarded from in-depth review due to high risk of bias ratings. The studies included in the full-text review phase are summarised in Table 3.6. Those above a 3* threshold are critically reviewed in substantial depth, and those above a 2* threshold are briefly reviewed in terms of their contribution to answering certain aspects of the research question.

Table 3.6. Studies that met criteria for updated review

Study	Topic	Context	Sample	Findings (including effect sizes, where given)	Trust-worthiness rating in this REA
Fleckenstein et al (2019)	Mathematic achievement in CLIL	German secondary schools	590 immersion students	CLIL students revealed a stronger increase from the first to the fourth grade in mathematics achievement than students in conventional programs, even when taking into consideration some initial advantages. No effect sizes reported and not possible to calculate.	3*
Fung & Yip (2014)	Academic achievement and motivation in physics	Hong Kong secondary schools	199 year-10 students	Learning in L1 is more beneficial for low-ability students, but L2 English is a more effective medium for high- achieving senior students due to motivating effects (Effect size of Cohen's d = 1.7 reported for ability).	3*
Isidro & Lasagabaster (2019)	CLIL and English competence, and content knowledge	Spanish secondary schools	44 students	The CLIL group made significantly greater gains in English competence, with no detrimental effect on the acquisition of content knowledge. No effect sizes provided, not possible to calculate.	2*
Lo & Murphy (2010)	Vocabulary knowledge	Hong Kong secondary schools	144 students	EMI students developed their vocabulary knowledge more than EFL students, thus the immersion environment provided a more favourable context for L2 vocabulary learning probably due to exposure to a greater number and variety of texts. No effect sizes provided.	2*
Merino & Lasagabaster (2018)	CLIL intensity and language development	Secondary schools in Spanish autonomous regions	393 students	Strong CLIL forms of education showed significantly better development in English language proficiency than weaker forms of CLIL (with a small effect size of r = 0.27). The study concludes that the more intensive the exposure to CLIL lessons, the better students' language develops	4*
Meyerhöffer & Dreesmann (2019a)	Academic achievement in biology	German secondary schools	158 students	The EMI students scored significantly better than the L1 medium students on biology tests with a small effect size (r = 0.26).	3*

Meyerhöffer & Dreesmann (2019b)	CLIL and content knowledge	German secondary schools	141 students	Post-tests revealed that the CLIL group scored higher on content knowledge after a teaching of a unit, although there may be selection bias. In this context, CLIL classes learned content knowledge as well as if not better than those in non-CLIL. (r = .29)	2*
Parvin et al. (2019)	CLIL and aviation English	Aviation programme in Iran	40 students	Language measures showed significant gains in English language knowledge in the randomly-assigned CLIL group with a large effect size (r = 0.80). The authors propose that CLIL positively affects attitudes to language learning.	2*
Perez Canado (2018)	CLIL and verbal intelligence, motivation and English language skills	Primary and secondary schools in Spanish autonomous regions	2,024 students	For primary education, statistically significant differences emerged on all the linguistic components and skills sampled, in favour of the CLIL group. Effect sizes, were low for listening, reading, and use of English (and disappeared when type of school was considered), but medium (d = .86) for productive speaking skills. After four additional years of CLIL, differences were reinforced in favour of CLIL with large effect sizes for all the linguistic skills.	3*
Salvador-García et al (2019)	Performance in CLIL Physical education	Spanish secondary schools	49 students	CLIL classes showed significantly greater physical activity with a large effect size (η2 = .332) than the members of the non-CLIL group, attributed to greater student attention in the CLIL group and more communicative teaching strategies.	2*

3.5.3 *Update of Lo and Lo (2014)*

Our risk of bias assessments of the four additional studies updated from Lo and Lo (2014) revealed one particular study that offered high quality research evidence: Fung and Yip (2014). Fung and Yip (2018) is discussed at length due to its focus on educational achievement in content learning, and because it was flagged in both the Lo and Lo (2014) and Graham et al (2018) as being of low risk of bias—then rated 3* overall by two independent assessors during updates of both seed reviews. Lo and Murphy (2010) received a score above the 2* threshold, and is briefly reviewed. A third study, Lau and Yuen (2011) met the inclusion criteria, but was ultimately discarded from in-depth review due to a high risk of bias due to questionable reliability of some components of the tests and a lack of clarity of the extent to which the comparison groups were well matched; there was little data on contextual information, school characteristics, or implementation. A fourth study, Hennebry and Gao (2018), met initial inclusion criteria and was rated as a high-quality study according to the sieve, but was later excluded due to its cross-sectional exploration of medium of instruction effects on motivation, rather than focusing on educational outcomes in an invention-style study.

Fung and Yip (2014) used a quasi-experimental design to compare academic achievement and academic motivation of students taking the same physics module in EMI and CMI. The students self-allocated to these groups, rather than being randomly allocated. A total of 199 year-10 students participated in three separate academic years (65, 66, and 68 in each year). The sample included 119 (c. 60%) boys and 81 (c. 40%) girls, but a Chi-square non-parametric tests revealed no significant differences between the CMI and EMI students in the distributions of sex (x2(1,199)1⁄4.019, p>.05). All participants attended the same CMI school, which primarily caters to medium-achieving students from working-class and poorer families. Chinese was the home language of all students and the parents mostly had relatively low levels of education (49% completed junior high school and 38% completed senior secondary school). A Chi-square non-parametric tests revealed no significant differences between the CMI and EMI students in terms of “ethnicity (x2(2,199)1⁄41.34, p>.05), family education level (x2(4,199)1⁄4.45, p>.05), or socioeconomic status (x2(1,199)1⁄4.019, p>.05), indicating that these factors had been statistically controlled prior to study commencement” (p. 1226). The students had previous contact with English primarily through mandatory English lessons, as well as through some out-of-class activities offered at the school.

To test for the influence of prior physics achievement, the authors collected all students’ previous (Year 9) exam results and evenly divided the students within each MOI into three achievement bands (low-, medium-, and high-achieving, modelled after the HK school internal assessments used to allocate places in secondary schools). Details are provided in the study on efforts to match both content and pedagogical tasks and approaches in the CMI and EMI courses. The classes were taught by the same instructor and held in an identical classroom setting, minimizing teacher effects. At the end of the physics course, all students took the same 4-hour examination in their medium of instruction, containing multiple-choice questions and structured problems. Students also completed pre- and post-tests assessing their conceptual knowledge of the covered topics in their respective mediums of instruction. These were supplemented with follow-up interviews with six randomly selected students from each medium of instruction group and year.

Analysis was conducted via a 2x3 factorial design: Medium of instruction (EMI and CMI) and Ability Group (low, medium, and high), with the dependent variables of physics achievement, while controlling for prior achievement (9th grade examinations). For the physics examination, the "results of factorial ANCOVA revealed significant main effects for both MOI (F(1, 192)1⁄44.51, p<.05) and Ability Group (F(2, 192)1⁄46.42, p<.005), as well as a significant interaction effect between the two (F(2, 192) 1⁄4 78.42, p < .001)." (p. 1129). Bonferroni-corrected post-hoc analyses showed significant differences between the low- and high-achieving EMI and CMI students (respectively p<.001, d=not reported and p<.005, d=1.7), but not the medium-achieving groups. Additional analysis was conducted on pre- and posts tests of conceptual knowledge and revealed mixed results: on one of the tests, there was no main effect found for medium of instruction, but follow-up tests indicated that the CMI students in the low-ability group had greatest improvement and the EMI students in the high-ability group made significantly greater improvement than their CMI peers. On the other test there was an overall main effect of medium of instruction (F(1, 193) 1⁄4 74.75, p < .001) and follow-up tests revealed the EMI students in the medium- and high-ability groups achieved significantly greater improvement than their CMI peers in the same groups. The authors conclude that the study finds evidence that learning in one's native language is beneficial for low-ability students at the senior secondary level in terms of making greater improvement in physics than learning in the L2, and that English is a more effective medium for high-achieving senior students due to motivating effects.

Lo and Murphy (2010) compared vocabulary knowledge of students receiving different modes of learning (primarily EMI vs. CMI instruction) and grade levels (year seven vs. year nine). Students from two secondary schools in the same district in Hong Kong were used, which the authors argued to be closely matched in terms of student ability, SES, previous English-learning experiences and out-of-school exposure to English. There was a total sample of 144 students after attrition. Data were collected via a receptive vocabulary test, a controlled productive vocabulary test, and a free productive vocabulary task. The study found significant differences between EMI and CMI—in all tests, the EMI students out-performed the CMI students. An ANCOVA was conducted to compare the grade nine scores while controlling for grade seven scores, and these differences remained statistically significant (p<.001), although it was unclear whether any differences were due to the programmes or existed a priori, as the scores were obtained from different groups of students who were already nine months into their programs of study. In interpreting these results, the researchers state that EMI students are exposed to English vocabulary in a wider variety and greater quantity of texts due to it being the medium of instruction for about 70% of their curriculum content, thus providing opportunities for vocabulary learning and use. They conclude that it "seems quite clear from this study, and others like it, that IM [EMI was designated as immersion in this study] provide a more favourable context for L2 vocabulary learning." (p. 234).

These two studies shed new light on the initial findings of Lo and Lo (2014) which concluded mixed findings of medium of instruction effects of content and language and suggested that EMI students might be sacrificing content learning for language gains. The studies also shed light on Lo and Lo's (2014) questioning of whether switching from L1 medium of instruction to EMI at grade 7 is too early in terms of students' L2 proficiency. Fung and Yip (2014) unravel the complexities of medium of instruction effects on students performing at different academic levels—highlighting that while low performing students benefit from L1 instruction, medium and high-performing students do not sacrifice content learning—at least not in the classes in their study. Lo and Murphy (2010) add some evidence that—at least in the realm of vocabulary—significant language learning improvements can be seen from switching medium of instruction at Year 7 in this context.

3.5.4 Update of Graham et al. (2018)

Our update of Graham et al (2018) revealed eight studies for full-text review. Our risk of bias evaluation indicated that Meyerhöffer and Dreesmann (2019a), Fung and Yip (2014) [already reviewed in the Lo and Lo 2014 update] and Fleckenstein et al (2019) were rated as 3* papers, and thus represented most rigorous research in terms of design and reporting. Four further studies (Salvador-Garcia et al. 2019; Parvin et al. 2019; Meyerhöffer & Dreesmann 2019b; Isidro & Lasagabaster 2019) attracted a 2* overall rating, and are reviewed in brief. One further study (Kuzminska et al. 2019) met our inclusion criteria but was discarded after full-text review due to a high risk of bias in the research design, intervention, and reporting.

Another Pladavell-Bellister (2019) also met inclusion criteria but was later discarded due to a focus on motivation rather than educational outcomes.

Meyerhöffer and Dreesmann (2019a) used a quasi-experimental design with in-tact classes of ninth grade German biology students, who were taking courses in English for first time. In total, six classes of Year 9 students were used from three different German secondary schools. Three of these were in the treatment group in the study (n=158). To ensure comparability of non-random groups, all students completed a set of tests to assess their psychological and cognitive preconditions. Students' average age was 14 years (range 13-16). There were fewer girls with just 40 (of 85) in the treatment group and 31 (of 73) in the control group. A bilingual teaching unit on immunology was developed and implemented as part of a project in the schools' curriculum. Several lessons were developed. A knowledge test consisted of eight close-ended items (five true/false, one multiple choice, two matching items) and eight short-answer items. All tasks were designed based on the curricular standards of the state, and the participating teachers confirmed their content being in accordance with the outcome requirements of their biology courses. Other measures of motivation were also taken.

The experimental group (Mdn = 7.00 out of 24 possible points, SD = 3.16) scored significantly better than the comparison group (Mdn = 5.50, SD = 2.42; U = 3904, p < .011, r = .26) as derived via a Mann–Whitney-U test, however both groups made significant gains from pre-test to post-test, indicating both models of education improved content knowledge. Researchers observed that students seemed to be motivated by the bilingual course, even though they were not self-selected into the group. The experimental group also had higher pre-test scores meaning that these may have been sustained to the post-test, although statistical tests controlling for this were not conducted on the non-normally distributed data. The study was limited to formal assessment that is typical for school contexts, and could not measure long-term content knowledge retention, nor the development of students' practical science skills. The authors state that the results of this study disprove concerns about reduced content knowledge gains of students taught through a second language, compared to students that learn the same topic in their mother tongue. While the study does offer some evidence for this claim, the lack of sensitivity to various factors influencing learning might have affected the strength of this claim.

Fleckenstein et al (2019) investigated mathematics achievement in one-way immersion over four years of elementary school. The design was a longitudinal study of matched groups, which were not randomly allocated, in five schools across two states in Germany, with 20 immersion classes in total (medium of instruction was English). Participants were 590 immersion/CLIL students (51.7% immersion). The mean age at the start of the study was 6 years. These classes were compared to non-CLIL students, where education was in German, but no fidelity to condition is provided. This 'big picture' study of L2 versus L1 mathematics teaching, used a number of standardised tests to measure outcomes, including the DEMAT (Deutscher Mathematiktest), however the test language was German for both groups, which may have disadvantaged the CLIL classes. Cognitive abilities were also tested to check for group differences. Mathematics performance was measured four times in grades in each year from Year One to Year Four, and a latent growth curve model was used. The initial level of mathematical achievement was higher for the CLIL group than for conventional programmes, hence there was bias in favour of CLIL programmes.

CLIL students revealed a stronger increase from the first to the fourth grade in mathematics achievement than students in conventional programs. The fact that content learned in an L2 was assessed in their L1 did not reverse these positive effects. The researchers suggest that use of an L2 for instruction may have positive effects on cognitive functions, supporting the idea that students have to 'work harder' because of the linguistic challenges associated with learning in an L2. However, a lack of fidelity to condition in teaching might have meant that more interesting teaching materials and techniques may have been used in the CLIL programme. The researchers report that CLIL students in general have higher initial levels of cognitive abilities and usually higher SES, but downplayed the possibility that higher SES families provide more academic support for their children. The study concludes that students studying mathematics via CLIL achieve better maths results even when taking into account initial advantages. While some evidence to support this is presented in the study, the degree to which the classes were truly matched in terms of SES, initial mathematics ability, and cognitive abilities, weakens the conclusions drawn.

Salvador-García, et al (2019) explored bilingual physical education to investigate the effects of CLIL on physical activity levels. The study used a quasi-experimental design of two randomly selected classes, although individuals within these groups were not randomly allocated. One group was taught physical education through CLIL (English) and one through the L1 (Spanish). There were attempts to match level of education, teacher and curriculum content. The CLIL class consisted of 23 students and the comparison had 26, with a gender balance in each. The intervention consisted of eight 50-minute sessions, six of which were practical and two theoretical. Each student had an accelerometer for the practical sessions to measure levels of physical activity. Mixed ANOVA was used to analyse the effect of the group variable on physical activity and showed significant effects for the group variable (F (1,47) = 23.38, p < .001, η2 = .332) in favour of the CLIL class who spent significantly more time in the active zones than the members of the non-CLIL group. Fidelity was high as the teaching was observed. In the interpretation of these results, the study suggests greater student attention in CLIL group and more communicative teaching strategies led to students paying more attention to explanations and thus increased participation. The small sample size and one-shot experiment, however, make it difficult to draw concrete conclusions.

Isidro and Lasagabaster (2019) explored the impact of CLIL on literacy development via a longitudinal comparison between CLIL (English L2) and non-CLIL groups (L1) studying social science. The small sample was taken from rural area of Spain made up of two groups of 3rd and 4th year secondary school: 44 students in total with 20 CLIL and 24 non-CLIL students. Students were not randomly assigned, but asked to enrol on to one of two programmes on a first come first served basis. Two placement tests showed both groups' prior competence was comparable in terms of L2 proficiency and previous knowledge of social science. The same curriculum and pedagogical components were used with both groups. During the project English competence, L1 competence, and content knowledge were tested three times (but tests were in L1 for both cohorts). Both groups made significant progress in English competence but the CLIL group made greater gains (but the effect sizes not given for the non-parametric tests). The CLIL group also improved to a greater extent compared to the non-CLIL group in L1 competence, which is hard to explain other than through theories of additive bilingualism. The CLIL group did not have detrimental effects compared to non-CLIL group in terms of social science knowledge, with no developmental differences between the two groups. No SES measurement was taken of two cohorts and it was an ‘opt-in’ study, which could be a factor in influencing these results.

In a completely different age group Parvin et al (2019) investigated trainee pilots’ learning of ‘aviation English’ through CLIL via a small-scale random control trial of 40 aviation students in Iran. Students were randomly allocated to a CLIL class with two teachers (content teacher and English teacher) and a non-CLIL (Farsi) class with one content teacher. Language measures showed no significant difference between the two groups at pre-test, but a significant difference at post-test with large effect size (r = 0.80). The authors draw on attitude measures taken during the study to propose that CLIL positively affects attitudes to language learning because the content becomes more meaningful. However, positive attitudes towards learning English are unsurprising, given these are future pilots who will need English in order to conduct their jobs safely. There were no measures of content knowledge, so the effect on other educational outcomes is unknown. The study also does not include a lot of information on ensuring fidelity of the intervention, such as what the non-CLIL class was like (e.g. studying aviation English in the L1 must have included some usage of L2 English, but this is not reported).

Meyerhöffer and Dreesmann (2019b) compared the teaching of a unit on immunology via CLIL to ‘bilingually inexperienced’ students to students studying biology in their L1 (German). The sample included six treatment classes of a total of 168 students and three comparison classes of a total of 73 students, all in Grade 9. All teachers received the same materials but had flexibility in delivery to mimic authentic conditions in most schools. The pre/post-test design included a content knowledge test, but it is unclear if this was standardized. Benchmark measures of individual differences revealed the CLIL group had significantly higher interest in biology (r=.29), but were not different in terms of self-regulation and self-efficacy. The post-test revealed that the CLIL group scored higher on content knowledge, but did not outperform controls to a significant level when the pre-test was accounted for in the analysis. The authors point to a selection bias in their sample of more gifted students choosing to take CLIL courses. Nevertheless, in this context, CLIL classes learned content knowledge as well as if not better than those in non-CLIL.

These studies do not change the original findings of Graham et al (2018) who found that there were some learning disadvantages for CLIL especially with 'low achieving students' (p. 32), and that many studies are subject to methodological problems due the fact that students (or their parents) elect to enrol in CLIL. Fung and Yip (2014) support the conclusion that low-performing students may be disadvantaged by a CLIL approach in terms of content learning. While Meyerhöffer and Dreesmann (2019a; 2019b) and Fleckenstein et al (2019) offer some evidence that CLIL students are not overall disadvantaged (to the contrary they claim positive learning effects), issues of self-selection have not been overcome in this new batch of studies, despite noteworthy efforts to match groups on a number of factors. Other studies in the update are simply too small in scale to offer any far-reaching implications.

3.5.5 Updates from Fitzpatrick et al. (2018)

In the update of Fitzpatrick et al (2018) two studies of relevance to the research question were found and judged to be of high quality and relevance. These were Merino and Lasagabaster (2018), rated as 4* overall, and Perez Canado (2018), rated as 3*. Both Merino and Lasagabaster (2018) and Perez Canado (2018) are reviewed in-depth in this section in terms of their contribution to answering the review question. Note that, Meyerhöffer and Dreesmann (2019b), which was reviewed as part of the update of Graham et al (2018), was also included in this update, and was rated as 2* by the reviewer.

Merino and Lasagabaster (2018) explored the effect of CLIL programmes' intensity on English proficiency, via a longitudinal design spanning one year that compared CLIL and non-CLIL taught groups. The sample in this study included students at lower SES levels in the Basque Autonomous Community (BAC) and in two neighbouring Spanish monolingual autonomous communities, Cantabria and La Rioja. Students were enrolled in Grade Seven (11–12 year olds) at T1, and in Grade Eight (12–13 year olds) at T2, and were balanced in gender representation. There were a total of 393 students who eventually took part at both stages of the study, after attrition between the two time points. The sample was comprised of three types of groups:

1. A ‘Non-CLIL’ group of 77 students in which Basque was the medium of instruction for all subjects except Spanish and English, comprised of students from eight high schools in the BAC.
2. A ‘CLIL-’ group of 208 students from the eight high schools in Cantabria and La Rioja, with an average of 3.4 CLIL sessions per week. These students had started receiving CLIL lessons during Grade Seven.
3. A ‘CLIL+’ group of 108 students from five high of the eight schools who had an average of 8.4 CLIL sessions per week. As with the CLIL- group, it was the first CLIL year for the students in this research group.

CLIL lessons covered different school subjects at the discretion of each high school, ranging from core subjects (maths, sciences or physical education), to elective ones offered in each school (arts and crafts or drama). English reading, writing, listening and the overall proficiency were assessed at T1 and T2 via a standardized Key English Test (KET). A speaking test was based on the story entitled “Frog, where are you?”—a wordless picture book in which participants were asked to describe what they were seeing in a series of pictures for a duration of three minutes. The speaking test was assessed by two raters who independently rated students' coherence (coefficient of intraclass correlation = 0.82; grammar = 0.75; fluency = 0.75; and pronunciation = 0.72). Competence in vocabulary was also assessed by means of a checklist which included key items and possible synonyms that the participants should have included in their productions.

A one-tailed analysis of variance (ANOVA) and two planned contrasts revealed significant differences in the overall proficiency in English according to the group. The first contrast (non-CLIL vs CLIL-) showed significant differences in favour of CLIL- (with a large effect size of r = 0.60). The second contrast (CLIL- vs CLIL+) revealed no significant differences between reduced CLIL exposure (CLIL-) and a higher exposure (CLIL+). A repeated measures ANOVA with two contrasts (Non-CLIL vs CLIL- and CLIL- vs CLIL+) explored language development between T1 and T2 and reflected a significant effect of time on proficiency in English. This means that all groups improved on language measures. Moreover, a statistically significant interaction was observed between the variables of time and group. This indicates that the evolution from T1 to T2 was different according to the group. No statistically significant differences were observed between the evolution of Non-CLIL and CLIL-. However, the contrast between CLIL- and CLIL+ showed that the evolution of CLIL+ students was significantly higher than that of their CLIL- counterparts (with a small effect size: r = 0.27).

There was a noticeable contrast between the cross-sectional results (T1) and the longitudinal ones (evolution from T1 to T2). When comparing the results at T1, both CLIL groups showed significantly higher scores than the non-CLIL group, but not between CLIL- and CLIL+. By contrast, when comparing the improvement of the three cohorts after a year, the CLIL-participants had progressed less than the CLIL+ participants, and at a similar amount to Non-CLIL when controlling for initial differences. The study concludes that the more intensive the exposure to CLIL lessons, the better students' language develops. That is, the effect of CLIL intensity is most evident for students who were participating in very intensive programmes. Similar gains in language development by Non-CLIL and CLIL- seem to indicate that CLIL will only produce a significant improvement across skill areas in the target language when it is part of a high intensity programme.

Perez Canado (2018) conducted a longitudinal experimental study comparing CLIL and EFL programmes on verbal intelligence, motivation and English language skills. The study drew on a sample of 2,024 students in 53 public, private, and charter schools in 12 provinces of three monolingual autonomous communities in Spain: Andalusia, Extremadura, and the Canary Islands. 828 students were finishing 6th grade of primary education (ages 11-12) and 1,196 were about to complete 4th grade of compulsory secondary education (ages 15-16). The majority of the cohort (78.3%) studied at public schools where CLIL branches and monolingual EFL streams co-exist. In turn, 17% of the pupils were enrolled in charter non-bilingual schools and the smallest percentage (4.7%) were private bilingual school students. 64% of the schools were located in urban areas, while the remaining 36% were rural. Practically equal percentages of schools delivered curriculum via partly CLIL streams (49%) and traditional EFL branches (51%) and there was a perfect balance in terms of gender (1,012 were male students and 1,011 were female). CLIL was positioned as the intervention, compared to EFL, but little detail was provided about the nature of CLIL delivery, nor the nature of the EFL classes. Further to this, as CLIL has a dual focus on content and language, it is unclear whether these two approaches to education are truly comparable. Verbal Intelligence, motivation and background information on the students was collected to ensure groups were matched, and representative of the larger student population. English Language Tests looking at vocabulary, reading, writing and speaking were used to explore language gains, which were designed and validated for this study. For primary education, statistically significant differences emerged on all the linguistic components and skills sampled, invariably in favour of the bilingual group. Effect sizes, however, were low for listening, reading, and use of English. Differences between the experimental and control groups were particularly marked for the productive speaking skills at the end of primary education, when CLIL students already outstripped their EFL counterparts on all the linguistic aspects sampled. After four additional years of participation in CLIL programs, the differences in English language competence were further reinforced, and were statistically significant in favour of the CLIL cohorts at extremely high confidence levels and with large effect sizes for all the linguistic aspects sampled. The researcher conclude that time is a crucial factor to ascertain the effects of CLIL on foreign language attainment, and the longer the students have been receiving bilingual education, the greater the differences are compared to their non-bilingual counterparts.

These studies add to the conclusions of Fitzpatrick et al (2018) that most work in this area continues to emerge from the Spanish context. While the Merino and Lasagabaster (2018) does not address issues noted in Fitzpatrick et al (2018) of self-selection, its incorporation of CLIL and non-CLIL investigations in lower SES school contexts does a lot to address the observed lack of research which addresses SES differences.

3.5.6 Conclusions

Despite the limitations of the evidence gathered in these reviews (and our updates of them), overall, we would support a number of key implications for L2 teachers and policymakers which arise from their work. These are summarized in Table 3.7.

Table 3.7. Summary of key implications for L2 teaching arising from our update of the medium of instruction reviews.

Topic / focus	Implication	Additional comments	Studies Contributing to this conclusion

CLIL/EMI and content learning	Teaching subjects in an L2 does not appear to harm content learning for high performing or self-selected students (but may do so for lower academic	If lower academic performers are in CLIL contexts, teachers should consider the facilitative role of some L1 use to ensure such students do not fall behind. Alternatively,	Binterová et al 2014;
			Dafouz et al 2014;
			Fleckenstein et al., 2019;
			Fung & Yip, 2014;
			Hernandez-Nanclares & Jimenez-Munoz, 2017;

	performers. Thus, institutional policy makers should consider detrimental effects on some students of only offering L2 medium classes	considerable and highly targeted language support should be offered to such students.	Meyerhöffer & Dreesmann, 2019a; 2019b
			Ouazizi, 2016;
			Piesche et al., 2016;
			Salvador-García, et al 2019;

CLIL/EMI and language development	At the school-level, teaching subjects in an L2 appears to have a positive effect on development of *some* L2 language skills compared to traditional language classes, with more consistent gains noted in receptive skills and vocabulary. Teachers could think about how to better harness the facilitative effects of immersion in their classes.	As CLIL/EMI may result in greater exposure to a wider variety of texts, leading to vocabulary development, teachers could think about how to support vocabulary development via increased use of authentic and modified texts.	Agustín-Llach & Canga Alonso, 2016;
			Agustín-Llach, 2016, 2017;
			Basterrechea & del Pilar García Mayo, 2014;
			Canga-Alonso, 2015a, 2015b;
			Dallinger et al., 2016;
			Gené-Gil et al 2015;
			Goris et al 2013;
			Ibarrola, 2012;
			Isidro & Lasagabaster, 2019;
			Jiménez et al 2009;
			Lazaro-Ibarrola, 2012;
			Lo & Lo, 2014
			Lo & Murphy, 2010;
			Lorenzo et al 2010;
			Manzano-Vázquez, 2014;
			Maxwell-Reid, 2010;
			Mesquida & Juan-Garau, 2013;
			Moore, 2011;
			Pérez Cañado, 2018;
			Pérez-Vidal & Roquet, 2015;
			Rallo-Fabra & Juan-Garau, 2011;
			Sylven 2010;
			Van der Leij et al. 2010;
			Xanthou, 2011;
			Yang, 2015

CLIL intensity	With little difference between weak CLIL programmes and traditional language classes in terms of language gains, school policy makers should not over-estimate the benefits of offering one or two CLIL subjects in a curriculum.	If schools decide to implement CLIL, they may see greater effects on language learning by offering a substantial part of the curriculum in the L2 (i.e. not 1-2 subjects).	Artieda et al 2017;
			Merino & Lasagabaster, 2018;
			Perez Canado, 2018;
			Ruiz de Zarobe, 2008

In drawing conclusions for the research question “What is the impact of using a non-native language as the medium of instruction in academic subjects on students’ academic outcomes?" and its subsidiary question "Are there implementation factors that lead to appositive impact on attainment of using a non-native language as the medium of instruction?" we conclude the following:

There is fairly convincing evidence that retaining the use of the L1 as the medium of instruction (in foreign language settings) might be more beneficial for low performing students (whether this low performance is due to general cognitive ability, language proficiency or both). However we are aware that this may lead to the further strengthening of socio-economic inequalities given the status of English as an international language.

There is a large body of somewhat weak evidence that suggests that CLIL does not harm content learning for most students in the long term

There is a body of somewhat convincing evidence that CLIL can improve content learning for high performing or gifted students due to motivational effects and to their possible initial advantages in L2 proficiency

Evidence exists that students who self-select (or are selected) into CLIL/EMI programmes tend to learn content just as well, if not better, than traditional programmes, and see greater gains in language development, especially in vocabulary and speaking, although this may not be due to medium of instruction effects alone, but due to the possibility that a) they start from a higher language proficiency level and/or b) their CLIL/EMI classes are in addition to EFL classes.

These conclusions need to be set against a methodological backdrop that retains a number of difficulties. We wish therefore to restate a number of issues which should be addressed by future research in this area.

Future research designs need to find ways of controlling for the selection process that often features in EMI/CLIL versus L1 MOI studies. As outlined earlier, Graham et al (2018) observe that the self-selection of students on programmes raises the possibility that they have greater motivation, or greater initial language proficiency or both. Lo and Lo (2014) argue that research needs to account for prior language proficiency, selection and enrolment when attempting a broad assessment of the effectiveness of EMI and CLIL. Goris et al (2019) argue that although many studies have shown benefits for CLIL, these are frequently affected by the attraction to CLIL or the selection of high-achieving or highly motivated students. Fleckenstein (2019) points to fact that CLIL students in general have higher initial levels of cognitive abilities and usually higher SES. While updated studies such as Fung and Yip (2014) included self-allocation, the researchers made great efforts to counterbalance baseline differences in groups, and to control for these in analysis. Meyerhöffer and Dreesmann (2019a) was also one of the few studies to randomise treatment at the (intact) class level, which somewhat strengthened their findings. Future research would benefit from a large-scale, multi-site study randomised control trial to circumvent the issue of self-selection. If this is not possible (as is the case for much educational research) Fung and Yip (2014) and Meyerhöffer and Dreesmann (2019a) offer some methodological procedures to try to minimise these confounding factors in analysis.

A further issue in answering this research question is to consider for what type of student CLIL is most beneficial. Graham et al (2018) found that there were some disadvantages of learning through the L2 for low achieving students. Fung and Yip (2014) was one of the few updated studies to look at low, medium and high performing students and found that the CMI students in the low-ability group and EMI students in the high-ability group showed the greatest improvement. Meyerhöffer and Dreesmann (2019b) looked at gifted students in CLIL classes, demonstrating that they learned content knowledge as well as, if not better than, those in non-CLIL, with no evidence of content knowledge deficit. While the evidence of CLIL benefits for academically high performing students is somewhat convincing, more research is clearly needed to unpack the costs and benefits of learning through an L2 on different types of students, as research shows the effects to vary according to student. With a research field saturated with studies on high-performing students, we particularly need to better understand the effect of L2 medium instruction on students who are lower academic performers and the linguistic support they will need in order to thrive in EMI/CLIL classrooms. More research is in general needed on students in lower SES contexts, although Merino and Lasagabaster (2018) and Fund and Yip (2014) have done some groundwork here.

Future research also needs to provide a better understanding of the effects of L2 instruction in more diverse contexts. Fitzpatrick observed many of the studies in their original review were conducted in Spain—a claim substantiated in this update. While the large number of studies which have been carried out in Spain and Hong Kong in particular provide some advantages in terms of uniformity, our overall ability to generalize beyond these geographical settings in limited. Particularly absent are investigations of the effects of L2 medium of instruction in Anglophone contexts, where languages other than English are used. Currently, we are lacking information on student uptake and the educational impact that a CLIL approach has when using a language that is not a global language.

Additionally, whether comparisons between CLIL and EFL are valid in terms of measuring language gains needs to be more closely scrutinised. Lo and Lo (2014) found that EMI students in their studies outperformed the CMI students in English language achievement. However, Fitzpatrick et al (2018) comment that "it is difficult to draw any strong conclusions in terms of the direct effect of CLIL as in most studies the amount of exposure to language was greater than for the control non-CLIL groups" (p. 60). As Lo and Murphy (2010) concluded in their study, students learning through L2 English are naturally exposed to English vocabulary in a wider variety and greater quantity of texts due to it being the medium of instruction, and thus this approach naturally provides a more favourable context for L2 vocabulary learning. While language learning benefits seem to be clear in CLIL contexts, where extensive exposure to the L2 is clearly offered, future research should try to investigate differences in intervention to ensure fidelity across contexts, if true comparisons are to be drawn. Merino and Lasagabaster (2018) was one of the few studies to explore intensity of CLIL, concluding that a small amount of CLIL was not better than traditional EFL. But a large amount of CLIL, if it is done right, was found to have very positive effects for learners' language development. This finding seems to indicate that research into the effects of different types of interventions is needed to better understand the language learning benefits of CLIL compared to EFL.

Future research should also more clearly specify the pedagogy that is being used both in the EMI/CLIL classrooms and the EFL/MFL classrooms that they are being compared with. We have already highlighted the potential diversity of pedagogy in both these sets of classrooms and without knowing "what is being taught and how" it is difficult to be confident with the outcome measures being adopted.

Whether observed benefits afforded by CLIL are maintained across subject areas also warrants further attention. Lo and Lo (2014) found that EMI students experienced a deficit in academic performance except in mathematics, however, this finding was not supported in any of our updated studies, partly because the comparison of academic subjects has rarely been a principal research question. When content learning was measured the studies in our review, these tended to explore L2 effects on content learning of mathematics and the hard sciences, with the exception of some very small-scale studies on social sciences and physical education. More attention needs to be paid to medium of instruction effects in different academic disciplines where, although language will be important in all of them, different aspects of language will feature quite strongly (e.g. technical language in science and mathematics).

Finally, future research needs to test the hypothesis that L1 medium of instruction benefits short term content learning and that EMI learners might catch up in the long term (see Marsh et al., 2002). A few of the longitudinal studies indicated time was an important factor to understanding the long-term effects of using the L2 as a medium of instruction. Fleckenstein et al (2019), for example, investigated mathematics achievement over four years of elementary school, but was one of the few studies to operate over this length of time. Future studies need to explore the long-term impact of medium of instruction on educational outcomes.

4. Synthesis of Findings

In this chapter we take the findings from each chapter as a whole in developing a cogent response to address each of the review questions (RQs). In so doing we will first summarise the main findings following from our searches and updates, and then articulate our conclusions regarding each review question.

4.1 Effective approaches to teaching foreign languages

The first review questions asks what approaches to teaching foreign languages have been adopted and what evidence is available as to their effectiveness. We considered this question together with the third review question which asks about practitioner skills and programme characteristics which also influence effective language teaching/learning in classroom-based settings. In addressing these two questions, we located two relatively recent reviews in Fitzpatrick, Clark, Tanguay and Tovey (2018) and Harris and O'Duibhir (2011), both of which addressed aspects of these two review questions.

Many volumes have been written describing different approaches to teaching foreign languages (e.g., Lopes & Cecilia (2019); Macaro, Graham & Woore, (2015); Richards & Rogers, (2014); Torres-Zúñiga & Schmidt (2017)). Our updates to Fitzpatrick et al (2018) and Harris and O'Duibhir (2011) did not reveal any new insights into which particular methods are more or less effective, nor any revelations about new methodologies previously unknown. Rather, our updates to these two seed reviews indicated that the more relevant question is really the third review question which concerns practitioner skills and programme characteristics. For example, Graham et al (2017) demonstrated that teacher experience was more predictive of FL success than the actual approach taken in comparing oral vs. literacy-based approaches. The effectiveness of different Focus on Form (FonF) approaches seems to depend on a variety of variables (e.g., L1 of learners, rule complexity). FonF approaches are effective when strategically employed and might be more beneficial for older learners. Indeed, the findings regarding an integration of meaning oriented approaches with a strategic focus on form does suggest that a pure focus on linguistic form may not be an evidentially-based approach. Our reviews also indicated that data-driven approaches may also be helpful regarding developing some key skills such as inductive reasoning approaches and metalinguistic awareness and could support the development of more autonomous learning. The research reviewed in Section 3.1 also indicates a role for technology and many programmes around the world make use of different learning technologies for teaching and learning FLs. The use of videos/television/films and their associated captions also seem to have a place in successful FL programmes. However, using technology requires careful thought and consideration and should not just be implemented without a clear understanding of how and what it will support in the FL classroom. Our review also demonstrated that the linguistic feature and/or communicative/language skill that is being targeted also is a primary influence on the effectiveness of any effective FL programme. Consequently, we review these in turn below.

The summary of research on FL programmes and vocabulary indicates that a key issue is the area of focus – the vocabulary item itself or on extracting meaning from text or discourse. Vocabulary can be learned in both more intentional and incidental conditions. Some form of direct teaching in combination with experience of the vocabulary item in context is beneficial (Hennebry et al, 2017). Furthermore, a high involvement load, where students are engaged and motivated can equally lead to higher gains in FL vocabulary learning. There is mixed evidence for imagery and songs. This is an important area for further research as this approach to teaching vocabulary is frequently used with young learners but apparently without much empirical evidence to demonstrate its effectiveness.

Our review of the evidence shows that explicit instruction of grammar is effective but other (less explicit) methods are equally so – and effectiveness of grammar instruction is influenced by proficiency where an explicit focus on grammar is perhaps less effective for beginning learners. To help younger learners develop grammatical competence, a focus on prosodic characteristics can be useful (Campfield & Murphy, 2017) which is associated with the finding that both oracy-based approaches (very common for younger learners) and literacy-based can be effective for developing grammatical knowledge. As with vocabulary, proficiency is a key variable – both in oral language and literacy skill. Processing instruction which provides structured input to learners can also be effective in developing grammatical competence. Interestingly while there was quite a substantial amount of research examining the relationship between technology and vocabulary learning, there was far less investigating the role of technological on grammatical development. This may be a profitable avenue for future work

Given reading consists of both lower-level (attending to letters/graphemes) and higher level (applying world/context knowledge) it is not surprising that approaches which support both can be helpful. Phonological training can help with lower-level processes (mapping phonemes to graphemes) but are less effective for semantics. Using technology appears again as a useful variable, as does videos/film/TV. Instruction of reading strategies can help learners develop comprehension and again, teacher competence in providing reading materials which reduce the cognitive load of the texts can also help support comprehension.

Literacy and oracy combinations can be effective, particularly where students are at different levels of competence, Technology again emerges as a potentially important variable but as with all instances of technology it matters how it is used. Similarly, as with reading, strategy training can be helpful with writing.

The findings related to developing speaking and listening skills suggests that meaningful interaction, particularly when there is FonF can be helpful. This finding reinforces the notion that FonF within meaning-based approaches are key. Game-based activities (not drill-based) can help with pronunciation. Technology (and specifically online interaction) can support development and afford learners opportunities. Affording learners the opportunity to make errors is good as this can lead to development. The use of authentic material has a place but as with the literature on reading, this needs to be carefully selected at the appropriate levels for students. Again, as with vocabulary and reading, a mixed method where students are engaged with more focused or targeted activities in addition to presentation of discourse can help. Finally, and also again as with vocabulary and reading, strategy instruction can be useful.

Our main findings of our update to Fitzpatrick et al (2018) highlight teacher proficiency in both the taught language and pedagogical knowledge; the use of rich methods and experience with language through a variety of media; good continuity between primary and secondary levels of education; both explicit and more implicit approaches; judicious use of technology; and strategy instruction.

The Harris and O'Duibhir (2011) review covers many of the same issues as Fitzpatrick et al (2018) and hence was also useful in addressing RQs 1 and 3. Whereas Fitzpatrick et al's (2018) review was carried out to speak to issues of language pedagogy in Wales, the Harris and O'Duibhir (2011) synthesis was commissioned by the Ireland's National Council for Curriculum and Assessment. Studies in this review (and subsequently updated in this REA) addressed questions about language teaching for ostensibly primary level learners in different language and education programmes. The five themes of the Harris and O'Duibhir (2011) review were centred around i) corrective feedback, ii) CLIL, iii) intensive language programmes (also figured in Fitzpatrick et al (2018); iv) communicative vs. analytical approaches; and v) literacy development in the L2. The main findings of the review, together with our updates illustrate that oral corrective feedback implemented in high quality oral interaction helps make learners become aware of their errors and helps develop appropriate reformulations. As with the research on FonF approaches reviewed in Fitzpatrick (2018) and relevant updates, FonF can be effective but needs to be carried out within the context of a rich, meaning-based communicative classroom. Finding the right balance is the key. Literacy instruction in primary languages (i.e., not just an oracy approach) has a place (a finding also manifest in Fitzpatrick et al (2018)). Many primary programmes eschew literacy-based skills development but there is credible evidence a combination of both oracy and literacy-based approaches can be useful for young learners. Also demonstrated in the Fitzpatrick et al (2018) review and updates, the Harris and O'Duibhir (2011) review provides empirical support for strategy instruction.

In summary, from the Fitzpatrick et al (2018) and Harris and O'Duibhir (2011) seed reviews, together with our updates for both, we see that no one-size-fits-all method works in relation to effective approaches. Rather, we need a skilled workforce (and one that is proficient in the language being taught), who can provide a rich, meaning-oriented context in the classroom, while at the same time drawing the learners' attention to linguistic form when necessary. A focus on both oral and literacy skills, even for younger learners is founded on empirical evidence. There is a place for appropriate use of technology if it actually enhances some aspect of language under pedagogical consideration. Similarly, using video/film/TV and captions can be supportive if used judiciously. Finally, both seed reviews and our respective updates have demonstrated the value of strategy instruction for foreign language learning.

4.2 The impact of foreign language learning on wider academic outcomes

The second review question asked what the impact of learning a foreign language is on wider academic outcomes. In implementing our methodology we were unable to find an appropriate seed review that had exclusively examined this question from the perspective of 'academic' outcomes – that is, school-based subjects and learning. However, we did find an appropriate two-part seed review that examined the question from a broader perspective. Fox, Corretjer, Webb and Tian (2019) and Fox, Corretjer and Web (2019) examined the research investigating the wider question of benefits of knowing more than one language. This included foreign language learning, bilingualism and multilingualism as well as academic achievement, cognitive abilities, attitudes and beliefs. Many of the studies discussed in these reviews did not adhere to the classic RCT or intervention-based design, however, did adhere to a classic experimental design (treatment + control group) where typically the grouping variable (bilingual vs not bilingual) was compared on some dependent measure(s). In Section 3.2 we provide a description of the main findings of Fox et al 1 and Fox et al 2 and discuss our updated reviews. Fox et al (1 and 2) group their reviews around specific themes, and we followed suit. Here we shall briefly summarise the main findings.

Cognitive abilities. Research in this area includes work on cognitive control, executive functioning, inhibitory control, working memory, attentional control, and cognitive flexibility. There is a great deal of research that has focused on examining whether bilinguals are advantaged in these areas. Much of the evidence in these areas suggests mixed support for the notion that knowing and using another language confers cognitive advantages. Given the studies vary in terms of samples (age of learner), level of bilingualism, specific tasks used, languages of the bilinguals, and so on, it is very difficult at this stage to determine whether and to what extent cognitive advantages in bilinguals can be reliably supported. While the question is inherently interesting from a psychological perspective, we would argue that there is sufficient evidence in other areas to demonstrate that knowing and using more than language has advantages – many of which are articulated in the Fox et al reviews and discussed in Section 3.2. As such, the mixed evidence concerning putative cognitive advantages for bilinguals is of nominal concern for educators.

Linguistic processing and reasoning. Research comparing monolinguals to bilinguals on their linguistic skills suggests that there is some evidence that bilinguals have enhanced skills in communicative competence. There is somewhat mixed evidence concerning vocabulary and reading skills – many studies in young, developing bilinguals for example suggest bilinguals have smaller vocabulary sizes within each of their two languages. However, other work suggests bilinguals have superior pragmatic skills. As with the cognitive advantages literature then, there is mixed evidence to suggest that being bilingual means that the learner is more adept at processing linguistic information.

Metalinguistic awareness. The research in this area is more unequivocal suggesting across numerous studies that bilinguals do have enhanced metalinguistic awareness, particularly for phonology, morphology and syntax. This is particularly relevant for educational contexts because research has clearly demonstrated that metalinguistic awareness predicts literacy development in children and one of the primary objectives of primary school is to develop literacy skills in students.

Cognitive development. As with the area of cognitive abilities we have mixed evidence concerning whether and to what extent bilingualism exerts an influence. Some research has demonstrated bilinguals have observably different cognitive architecture than monolinguals but more recent research has suggested otherwise. As with the work on cognitive abilities, we would argue that this work is less germane to an educational perspective.

Cross-language activation. Associated with the work on cognitive abilities, there is some credible evidence to suggest that bilinguals are skilled at language switching. However, this is hardly surprising in that monolinguals have but one language and hence switching across languages is an impossibility for them. A more interesting question is whether this skill transfers on to other skills, which has been less well-researched.

Spatial reasoning. There is some evidence suggesting that bilinguals have superior spatial reasoning skills. While this work was not carried out within an educational context, it could have applications to certain aspects of curriculum, such as maths for example.

Academic achievement. Within the broader area of academic achievement, the Fox et al (1 and 2) reviews, together with our updates, demonstrated that there was evidence that bilingual education and bilingualism might enjoy some advantages in the domains of language and literacy over monolinguals. However, there is more equivocal evidence for these advantages in the work on reading comprehension. Studies comparing bilinguals against monolinguals on maths achievement found no differences overall (despite the putative advantages for bilinguals on spatial reasoning as above). However, there was some evidence that bilinguals can make more rapid progress on maths. Therefore, while they end up at the same place as monolinguals in terms of achievement, they might get there more quickly. Other studies suggest clear advantages for bilinguals on maths, science and social studies.

There is also mixed evidence concerning the impact of bilingualism on **creativity** and **aging and health** along with more positive evidence suggesting bilinguals have superior **intercultural competence**, might be more **employable** and might be more **motivated** learners.

The evidence then concerning wider impacts of bilingualism/foreign language learning on other areas is both diverse and mixed. Many different areas have been researched, most notable cognitive abilities. What is missing from the research is a sustained and systematic programme of research which has examined the impact of learning foreign languages within school-based settings on other curricular subjects. We have some evidence to suggest that there are indeed positive impacts on FL learning on aspects of developing L1 literacy (Murphy et al, 2015) as previously indicated. We urgently need more carefully constructed, rigorous research in this area so as to be able to speak directly to this question. In the meantime, the Fox et al (1 and 2) reviews together with our updates do suggest that there are many dimensions which could be positively influenced by knowing and using another language and which could profitably be better understood by educators.

4.3 The impact of using a non-native language as medium of instruction on academic outcomes

In this section we consider the outcomes of our review of research on educational programmes where children are taught academic subjects through the medium of a non-native language. Many programmes have been developed where (academic) content is integrated with the learning of a language, variously known as immersion, CLIL, CBI or EMI (when the language of instruction is English). We also consider this research in light of the fifth review question asking about implementation factors of medium of instruction programmes that facilitate learning. In addressing these RQs we reviewed the research in four seed reviews (Graham et al 2018; Fitzpatrick et al, 2018; Lo & Lo, 2014; and Goris et al 2019) and following our methodology carried out updates to these reviews. In summary, there is evidence from each of these four reviews in favour of medium of instruction over traditional foreign language programmes. However, as with the other questions we have addressed in this REA, the evidence is mixed and all four seed reviews point to methodological issues/variability as potential explanations for the ambiguous evidence reported. We highlight in Section 3.4 that what is missing – which is true for all of the areas reviewed in this REA – is systematic, rigorous research which provides reliable and consistent answers to the questions posed in this review.

Our updates to the four seed reviews addressing RQs 4 and 5 suggest that student proficiency interacts with effectiveness of learning through the L1 or L2. This is a similar pattern observed in other questions in this review that the students' skills and proficiency levels is a key variable that interacts with outcomes. Similar to these findings, our update to Graham et al (2018) suggest there is some evidence of learning disadvantages for low-achieving students in CLIL programmes but again, as with the original seed review, there is mixed evidence here as Meyerhöffer and Dreesman (2019a; 2019b) and Fleckenstein et al (2019) both reported some evidence that CLIL students are advantaged in terms of learning outcomes. We also have evidence that EMI can help improve vocabulary growth as manifest in Lo and Murphy, (2010). We note in Section 3.3, however, that future research should endeavour to factor in a number of areas so as to improve the evidence base addressing these questions. These include controlling how participants are selected for inclusion in medium of instruction (MoI) studies, examining more closely the interaction between individual differences and learning outcomes in MoI contexts, a closer examination of the pedagogy within these contexts to better explicate what specific pedagogical approaches within MoI settings are more, or less, effective, more research examining more content areas and longer term benefits (or lack thereof) of learning through the medium of the non-native language.

4.4 Summary

While the questions posed in this REA are vast in scope and there is a considerable amount of research that speaks to these issues, there are some key findings that have emerged from our review. There is a considerable range of approaches being adopted world-wide in respect of language learning through educational provision. These range from highly input-limited instructed foreign language contexts where students receive less than one hour per week of class time being taught a FL to rich, communicatively oriented, and content-based input where the to-be-learned language is used as a medium of instruction (MoI).

Reviews of research has already demonstrated that immersion models tend to be more successful in developing higher levels of L2 knowledge than input-limited taught programmes (e.g., Murphy, 2014). The seed reviews and our updates therein have highlighted that while there are clearly manifest differences depending on the type of programme, what is more predictive of students' outcomes is not the programme per se but a host of other variables. These include teachers' pedagogical knowledge and skills, teachers' proficiency in the L2, students' individual differences (including proficiency in both L1 and L2 as well as other variables such as motivation and support in the home). It also matters what feature of language is being investigated (e.g., vocabulary vs grammar) and which language skill (reading, writing, speaking or listening) is being measured. Variability in outcomes can be accounted for by variability in methodological approaches taken by the different studies, teachers' level of knowledge and skill, and the kind of language being investigated. Our reviews also demonstrated that there can be a positive impact of the use of technology and video/film/tv in language in education contexts but that these need to be used judiciously. In other words, the teachers' pedagogical skill in using technology and for specific purposes influences whether and to what extent use of technology enhances and/or supports learning outcomes. In short, a one-size-fits-all approach is not appropriate here, but rather, the success of a given approach to supporting language development through education is multi-faceted and depends very much of the context, knowledge and skills of both teachers and students.

In addressing the third review question asking about impact of learning a foreign language on wider academic outcomes we found that there has been little systematic research on this area that speaks directly to academic outcomes in educational contexts. There has been considerable research investigating putative cognitive advantages for knowing and using more than one language, the evidence in this area is mixed and consequently at present inconclusive. Whether or not future work is able to put this issue to rest is somewhat irrelevant for the purposes of our REA. Knowing and using another language is advantageous, because it allows the individual to know and use another language. Circular reasoning such as this should normally be eschewed but we use it here to demonstrate a self-evident truth – being knowledgeable in another language is a good thing in and of its own right.

Arguably a more interesting question for this REA is research specifically examining whether learning a language within school settings has an impact on other academic content areas (e.g., maths, science, literacy). Unfortunately, our review has indicated these are few in number and we therefore would argue we urgently need more systematic research in this area. That which we have has offered some positive evidence that learning a FL in school can lead to positive outcomes in other areas but given the lack of research in this area this is only a tentative conclusion at this stage.

Finally, in addition to the variability inherent in the different studies, we also report throughout the review variability in the strength and rigour of the research itself. The studies reported in the Fitzpatrick et al (2018) seed review were all deemed of high quality. However, given our approach here was to adopt the same methods used in each seed review, necessarily not all research reviewed in the remaining seed reviews and our respective updates include only high-quality, rigorous research. We have discussed throughout our narrative when we felt studies were more or less convincing. However, the fact remains that there is a great deal of variability here and we could profitably spend future research developing a more consistently robust research agenda.

5. References

* Marks systematic reviews and meta-analyses longlisted during Phase 1.

** Marks publications selected as 'seed reviews' during Phase 1.

† Marks studies selected for updating the seed reviews during Phase 2.

Abdelrahman, L. A. M., Dewitt, D., Alias, N., & Rahman, M. N. A. (2017). Flipped learning for ESL writing in a Sudanese school. *Turkish Online Journal of Educational Technology, 16*(3), 60-70.

Adi-Japha, E., Berberich-Artzi, J., & Libnawi, A. (2010). Cognitive flexibility in drawings of bilingual children. *Child Development*, *81*, 1356–1366.

Alcón, E. (2007). Incidental focus on form, noticing and vocabulary learning in the EFL classroom. *International Journal of English Studies, 7*(2), 41-60.

Aldosari, A., & Alsultan, M. (2017). The influence of early bilingual education (English) on the first language (Arabic) literacy skills in the second grade of elementary school: Saudi Arabia. *Journal of Education and Practice*, *8*(5), 135–142.

*Alexander, L. (2019). *What do we know about the effectiveness of group work in Japan's secondary English education? A systematic review of an 'active learning' technique* (Master's dissertation). University of Oxford, Oxford, United Kingdom.

†Aljohani, O. (2016). Does teaching English in Saudi primary schools affect students' academic achievement in Arabic subjects? *Advances in Language and Literary Studies*, *7*(1), 214–225.

*Alsadhan, R. O. (2011). *Effect of textual enhancement and explicit rule presentation on the noticing and acquisition of L2 grammatical structures: A meta-analysis* (Master's dissertation). Colorado State University, Fort Collins, Colorado, USA.

†Alvarez-Marinelli, H., Blanco, M., Lara-Alecio, R., Irby, B. J., Tong, F., Stanley, K. & Fan, Y. (2016). Computer assisted English language learning in Costa Rican elementary schools: An experimental study. *Computer Assisted Language Learning*, *29*(1), 103–126.

Amer, A. (1997). The effect of the teacher's reading aloud on the reading comprehension of EFL students. *ELT Journal*, *51*(1), 43-47.

†Antón, E., Carreiras, M, Duñabeitia, J.A. (2019). The impact of bilingualism on executive functions and working memory in young adults. *PLoSONE*, *14*(2), e0206770.

†Arantzeta, M., Howard, D., Webster, J., Laka, I., Martínez-Zabaleta, M. & Bastiaanse, R. (2019). Bilingual aphasia: Assessing cross-linguistic asymmetries and bilingual advantage in sentence comprehension deficits. *Cortex*, *119*, 195–214.

Ardasheva, Y., Wang, Z., Adesope, O. & Valentine, J. (2017) 'Exploring Effectiveness and Moderators of Language Learning Strategy Instruction on Second Language and Self-Regulated Learning Outcomes'. *Review of Educational Research*, *87*(3): 544-582

Arredondo, M. M., Hu, X. -S., Satterfield, T., & Kovelman, I. (2017). Bilingualism alters children's frontal lobe functioning for attentional control. *Developmental Science*, *20*(3).

Astle, D. E. & Scerif, G. (2009). Using Developmental Cognitive Neuroscience to Study Behavioral and Attentional Control. *Developmental Psychobiology*, *51*(2): 107–118. doi:10.1002/dev.20350. PMID 18973175.

Babcock, L., & Vallesi, A. (2017). Are simultaneous interpreters expert bilinguals, unique bilinguals, or both? *Bilingualism*, *20*(2), 403–417.

Bak, T. H., Nissan, J. J., Allerhand, M. M., & Deary, I. J. (2014). Does bilingualism influence cognitive aging? *Annals of Neurology*, *75*(6), 959–963.

†Balcı, Ö. & Çakır, A. (2012). Teaching vocabulary through collocations in EFL classes: The

case of Turkey. *International Journal of Research Studies in Language Learning*, *1*(1), 21–32.

Barski, E., & Wilkerson-Barker, D. (2019). Making the most of general education foreign language requirements. *Foreign Language Annals*, *52*, 491–506.

Bartolotti, J., Bradley, K., Hernandez, A. E., & Marian, V. (2017). Neural signatures of second language learning and control. *Special Issue: The Neural Basis of Language Learning*, *98*, 130–138

Bauckham, I., 2016. *Modern foreign languages pedagogy review: A review of modern foreign languages teaching practice in key stage 3 and key stage 4*. Teaching Schools Council.

†Bavi, F. (2018). The effect of using fun activities on learning vocabulary at the elementary level. *Journal of Teaching and Research*, *9*(3), 629–639.

Beadle, S., Humburg, M., Smith, R., & Vale, P. (2016). Study on foreign language proficiency and employability: Executive summary 1. *European Journal of Language Policy*; Liverpool, *8*(2), 243–253.

Belpoliti, F., & Pérez, M. E. (2019). Service learning in Spanish for the health professions: Heritage language learners' competence in action. *Foreign Language Annals*, *52*, 529–550.

†Berens, M. S., Kovelman, I. & Petitto, L.-A. (2013). Should bilingual children learn reading in two languages at the same time or in sequence? *Bilingual Research Journal*, *36*(1), 35–60.

Bialystok, E., & Barac, R. (2012). Emerging bilingualism: Dissociating advantages for metalinguistic awareness and executive control. *Cognition*, *122*, 67–73.

Bialystok, E., & Feng, X. (2009). Language proficiency and executive control in proactive interference: Evidence from monolingual and bilingual children and adults. *Brain and Language*, *109*, 93–100.

Bialystok, E., & Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science*, *19*, 19–23.

Bien-Miller, L., Akbulut, M., Wildemann, A., & Reich, H. H. (2017). The relationship between bilingualism and metalinguistic awareness in primary school children. *Zeitschrift Fur Erziehungswissenschaft*, *20*(2), 193–211.

Blom, E., Boerma, T., Bosma, E., Cornips, L., & Everaert, E. (2017). Cognitive advantages of bilingual children in different sociolinguistic contexts. *Frontiers in Psychology*, *8*(552), 1–12.

Blom, E., Küntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: Working memory in bilingual Turkish–Dutch children. *Journal of Experimental Child Psychology*, *128*, 105–119.

Blumenfeld, H. K., & Marian, V. (2011). Bilingualism influences inhibitory control in auditory comprehension. *Cognition*, *118*, 245–257.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348-393.

Brumen, M. (2011). The perception of and motivation for foreign language learning in pre-school. *Early Child Development and Care*, *181*, 717–732.

*Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*(5), 603–632.

†Buckingham, L. & Alpaslan, R. S. (2017). Promoting speaking proficiency and willingness to communicate in Turkish young learners of English through asynchronous computer-mediated practice. System, *65*, 25–37.

Buga, R., Cāpeneatā, I., Chirasnel, C., & Popa, A. (2014). Facebook in foreign language teaching - a tool to improve communication competences. *Procedia - Social and Behavioral Sciences, 128*, 93-98

*Butler Stewart, K. (2019). *A meta-analysis of the relationship between learning a foreign language in elementary school and student achievement* (EdD dissertation). University of Alabama, Tuscaloosa, Alabama, USA.

Cable, C., Driscoll, P., Mitchell, R., Sing, S., Cremin, T., Earl, J., et al. (2010). *Languages learning at Key Stage 2: A longitudinal study*. Runcorn: Department for Children, Schools and Families.

Camo, A. C., & Ballester, E. P. (2015). The effects of using L1 translation on young learners' foreign language vocabulary learning. *Elia-Estudios De Linguistica Inglesa Aplicada, 15*(15), 109-134.

Campfield, D. & Murphy, V.A. (2017). The influence of prosodic input in the second language classroom: Does it stimulate child acquisition of word order and function words? *Language Learning Journal* http://dx.doi.org/10.1080/09571736.2013.807864

†Cañado, M. L. P. (2018). CLIL and educational level: A longitudinal study on the impact of CLIL on language outcomes. *Porta Linguarum*, *29*, 51–70.

Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science*, *11*, 282–298.

Chalmers H (2018). Why all the fuss about Randomised Trials? *ResearchED Magazine. 1(2)*,12-13.

†Chan, M. (2018). Processing instruction in helping map forms and meaning in second language acquisition of English simple past. *The Journal of Educational Research, 111*(6), 720-732.

†Chan, M. (2019). The role of classroom input: Processing instruction, traditional instruction, and implicit instruction in the acquisition of the English simple past by Cantonese ESL learners in Hong Kong. *System*, *80*, 246–256.

*Chang, M.-M. & Lin, M.-C. (2013). Strategy-oriented web-based English instruction – A meta-analysis. *Australasian Journal of Educational Technology*, *29*(2), 203–216.

Chen, C. M., Tan, C. C., & Lo, B. J. (2016). Facilitating English-language learners' oral reading fluency with digital pen technology. *Interactive Learning Environments, 24*(1), 96-118.

†Chen, Y.-R., Liu, Y.-T., Todd, A. G. (2018). Transient but effective? Captioning and adolescent EFL learners’ spoken vocabulary acquisition. *English Teaching & Learning*, *42*, 25–56.

Chertkow, H., Whitehead, V., Phillips, N., Wolfson, C., Atherton, J., & Bergman, H. (2010). Multilingualism (but not always bilingualism) delays the onset of Alzheimer disease: Evidence from a bilingual community. *Alzheimer Disease & Associated Disorders*, *24*, 118–125.

*Chiu, Y.-H. (2013). Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology*, *44*(2), E52–E56.

Choi, J. Y., Jeon, S., & Lippard, C. (2018). Dual language learning, inhibitory control, and math achievement in Head Start and kindergarten. *Early Childhood Research Quarterly*, *42*, 66–78.

Choi, J. Y., Rouse, H., & Ryu, D. (2018). Academic development of Head Start children: Role of dual language learning status. *Journal of Applied Developmental Psychology*, *56*, 52–66.

Chung-Fat-Yim, A., Sorge, G. B., & Bialystok, E. (2017). The relationship between bilingualism and selective attention in young adults: Evidence from an ambiguous figures task. *The Quarterly Journal of Experimental Psychology*, *70*(3), 366–372.

Claassen, J., Jama, Z., Manga, N., Lewis, M., & Hellenberg, D. (2017). Building freeways: Piloting communication skills in additional languages to health service personnel in Cape Town, South Africa. *BMC Health Services Research*, *17*(1), 390.

Cockcroft, K., Wigdorowitz, M., & Liversage, L. (2019). A multilingual advantage in the components of working memory. *Bilingualism*, *22*(1), 15–29.

Coelho, D., Andrade, A. I., & Portugal, G. (2018). The “Awakening to Languages” approach at preschool: Developing children's communicative competence. *Language Awareness*, *27*(3), 197–221.

*Cole, M. W. (2013). Rompiendo el silencio: Meta-analysis of the effectiveness of peer-mediated learning at improving language outcomes for ELLs. *Bilingual Research Journal*, *36*(2), 146–166.

*Cole, M. W. (2014). Speaking to Read: Meta-Analysis of Peer-Mediated Learning for English Language Learners. *Journal of Literacy Research*, *46*(3), 358–382.

Collins, B. A., Toppelberg, C. O., Suárez-Orozco, C., O’Connor, E., & Nieto-Castañon, A. (2011). Cross-sectional associations of Spanish and English competence and well-being in Latino children of immigrants in kindergarten. *International Journal of the Sociology of Language*, *1*, 5–23.

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*(4), 655-680.

Colzato, L. S., Bajo, M. T., van den Wildenberg, W., Paolieri, D., Nieuwenhuis, S., La Heij, W., & Hommel, B. (2008). How does bilingualism improve executive control? A comparison of active and reactive inhibition mechanisms. *Journal of Experimental Psychology-Learning Memory and Cognition*, *34*, 302–312.

†Comishen, K. J., Bialystok, E. & Adler, S. A. (2019). The impact of bilingual environments on selective attention in infancy. *Developmental Science*, *22*(4), e12797.

Connelly, P., Briggart, A., Miller, S., O’Hare, L. & Thurston, A. (2017). *Using randomised control trials in Education*. Sage

Cooper, T. C., Yanosky, D. J., Wisenbaker, J. M., Jahner, D., Webb, E., & Wilbur, M. L. (2008). Foreign language learning and SAT verbal scores revisited. *Foreign Language Annals*, *41*, 200–217.

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.

†Coyle, Y. & Roca de Larios, J. (2014). Exploring the role played by error correction and models on children’s reported noticing and output production in a L2 writing task. *Studies in Second Language Acquisition*, *36*(3), 451–485.

Craik, F. I. M., Bialystok, E., & Freedman, M. (2010). Delaying the onset of Alzheimer disease: Bilingualism as a form of cognitive reserve. *Neurology*, *75*, 1726–1729.

Critical Appraisal Skills Programme (2018). *CASP Systematic Review Checklist*. Retrieved 17 Sept 2019 from https://casp-uk.net/wp-content/uploads/2018/01/CASP-Systematic-Review-Checklist_2018.pdf

Daller, M., & Ongun, Z. (2018). The Threshold Hypothesis revisited: Bilingual lexical knowledge and non-verbal IQ development. *International Journal of Bilingualism*, *22*(6), 675–694.

Damari, R. R., Rivers, W. P., Brecht, R. D., Gardner, P., Pulupa, C., & Robinson, J. (2017). The demand for multilingual human capital in the U.S. labor market. *Foreign Language Annals, 50*, 13–37.

†Damian, M. F., Ye, W., Oh, M. & Yang, S. (2019). Bilinguals as "experts"? Comparing performance of mono- to bilingual individuals via a mousetracking paradigm. *Bilingualism: Language and Cognition*, *22*(5), 1176–1193.

Davis, G. M. (2017). Songs in the young learner classroom: A critical review of evidence. *ELT Journal, 71*(4), 445-455.

de Bruin, A., Treccani, B., & Sala, Della, S. (2014). Cognitive advantage in bilingualism. *Psychological Science*, *26*(1), 99–107. http://doi.org/10.1177/0956797614557866.

De Sousa, D. S. (2012). Literacy acquisition and bilingualism: The effect of biliteracy instruction on English reading achievement in bilingual Afrikaans-English South African children. *Journal of Communications Research*, *4*(4), 301–334.

Del Maschio, N., Sulpizio, S., Gallo, F., Fedeli, D., Weekes, B. S., & Abutalebi, J. (2018). Neuroplasticity across the lifespan and aging effects in bilinguals and monolinguals. *Brain Cognition*, *125*, 118-126. doi: 10.1016/j.bandc.2018.06.007.

Dewaele, J. (2010). Multilingualism and affordances: Variation in self-perceived communicative competence and communicative anxiety in French L1, L2, L3 and L4. *IRAL: International Review of Applied Linguistics in Language Teaching*, *48*, 105–129.

DfE (Department for Education) (2016) *Eliminating unnecessary workload around marking: Report of the Independent Teacher Workload Review Group*. Accessible at: www.gov.uk/government/publications/reducing-teacher-workload-marking-policy-review-group-report

Di Martino, E. & Di Sabato, B. (2012) CLIL implementation in Italian schools: Can long-serving teachers be retrained effectively? The Italian protagonists' voice. *Latin American journal of Content and Language Integrated Learning, 5*(2), 73-105.

Diaz, V., & Farrar, M. J. (2018). The missing explanation of the false-belief advantage in bilingual children: A longitudinal study. *Developmental Science*, *21*(4), 1–13.

Dillon, A. M. (2009). Metalinguistic awareness and evidence of cross-linguistic influence among bilingual learners in Irish primary schools. *Language Awareness*, *18*, 182–197.

Dolean, D. D. (2014). Using the keyword method in the classroom: Is the interacting imagery necessary? *System, 45*, 17-26.

Dolean, D. D., & Dolghi, A. (2016). Teaching young FL learners new vocabulary: A comparison between the efficiency of Keyword Method and Total Physical Response. *International Journal of English Linguistics, 6*(6), 1-7.

Domínguez, R., & Pessoa, S. (2005). Early versus late start in foreign language education: Documenting achievements. *Foreign Language Annals*, *38*, 473–483.

Drew, I. (2009). Using the Early Years Literacy Programme in primary EFL Norwegian classrooms. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 108-120). Bristol: Multilingual Matters.

*Driscoll, P., Jones, J., Martin, C., Graham-Matheson, L., Dismore, H. & Sykes, R. (2004). A systematic review of the characteristics of effective foreign language teaching to pupils between the ages of 7 and 11. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Duran, C. S. (2016). "I want to do things with languages": A male Karenni refugee's reconstructing multilingual capital. *Journal of Language, Identity & Education*, *15*(4), 216–229.

†Ebert, K. D., Rak, D., Slawny, C. M. & Fogg, L. (2019). Attention in bilingual children with developmental language disorder. *Journal of Language and Hearing Research*, *62*(4), 979–992.

Edelenbos, P., Johnstone, R., & Kubanek, A. (2006). *The main pedagogical principles underlying the teaching of languages to very young learners. Languages for the children of Europe: Published research, good practice and main principles*. Final report of the EAC 89/04, Lot 1 study European Commission Education and Culture, Retrieved from http://ec.europa.eu/education/languages/pdf/doc425_en.pdf

Elgün-Gündüz, Z., Akcan, S., Bayyurt, Y. (2012). Isolated form-focused instruction and integrated form-focused instruction in primary school English classrooms in Turkey. *Language, Culture and Curriculum, 25*(2), 157-171.

†Fecher, N. & Johnson, E. K. (2019). Bilingual infants excel at foreign-language talker recognition. *Developmental Science*, *22*(4), e12778.

Fernandez, M., Tartar, J. L., Padron, D., & Acosta, J. (2013). Neurophysiological marker of inhibition distinguishes language groups on a non-linguistic executive function test. *Brain and Cognition*, *83*(3), 330–336.

†Festman, J. & Schwieter, J. W. (2019). Self-concepts in reading and spelling among mono- and multilingual children: Extending the bilingual advantage. *Behavioral Sciences*, *9*(4), 39.

Fidaoui, D., Bahous, R., & Bacha, N. N. (2010). CALL in Lebanese elementary ESL writing classrooms. *Computer Assisted Language Learning, 23*(2), 151-168.

†Figueroa Murphy, A. (2014). The effect of dual-language and transitional-bilingual education instructional models on Spanish proficiency for English language learners. *Bilingual Research Journal*, *37*(2), 182–194.

**Fitzpatrick, T., Morris, S., Clark, T., Mitchell, R., Needs, J., Tanguay, E. & Tovey, B. (2018). *Rapid Evidence Assessment: Effective Second Language Teaching Approaches and Methods*. Report commissioned by the Welsh Government. Retrieved 26 Sep 2019 from: https://gov.wales/effective-second-language-teaching-approaches-and-methods-rapid-evidence-assessment

†Fleckenstein, J., Gebauer, S. K. & Möller, J. (2019). Promoting mathematics achievement in one-way immersion: Performance development over four years of elementary school. *Contemporary Educational Psychology*, *56*, 228–235.

Folke, T., Ouzia, J., Bright, P., De Martino, B., & Filippi, R. (2016). A bilingual disadvantage in metacognitive processing. *Cognition*, *150*, 119–132.

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.

**Fox, R., Corretjer, O., Webb, K. & Tian, J. (2019). Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2005–2011. *Foreign Language Annals*, *52*(3), 470–490.

**Fox, R., Corretjer, O. & Webb, K. (2019). Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2012–2019. *Foreign Language Annals*, *52*(4), 699–726.

†Fung, D. & Yip, V. (2014). The effects of medium of instruction in certificate-level Physics on achievement and motivation to learn. *Journal of Research in Science Teaching*, *51*(10), 1219–1245.

Fürst, G., & Grin, F. (2018). Multilingualism and creativity: A multivariate approach. *Journal of Multilingual and Multicultural Development*, *39*(4), 341–355.

Garrett, R. S. (2011). *Multilingualism, mathematics achievement and instructional language policy*. Dissertation Abstracts International: Section A. Humanities and Social Sciences, 3523.

Ghonsooly, B., & Showqi, S. (2012). The effects of foreign language learning on creativity. *English Language Teaching*, *5*(4), 161–167.

Godoy, R., Reyes-Garcia, V., Seyfried, C., Huanca, T., Leonard, W. R., McDade, T., Tanner, S., & Vadez, V. (2009). Language skills and earnings: Evidence from a pre-industrial economy in the Bolivian Amazon. *Economics of Education Review*, *26*, 349–360.

Gogonas, N., & Kirsch, C. (2018). "In this country my children are learning two of the most important languages in Europe": Ideologies of language as a commodity among Greek migrant families in Luxembourg. *International Journal of Bilingual Education and Bilingualism*, *21*(4), 426–438.

†Gonzalez-Barrero, A. M. & Nadig, A. S. (2019). Can bilingualism mitigate set-shifting difficulties in children with autism spectrum disorders? *Child Development*, *90*(4), 1043–1060.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, *110*, 47–59.

**Goris, J. A., Denessen, E. J. P. G. & Verhoeven, L. T. W. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies. *European Educational Research Journal*, *18*(6), 675–698.

**Graham, K. M., Choi, Y., Davoodi, A., Razmeh, S. & Dixon, L. Q. (2018). Language and content outcomes of CLIL and EMI: A systematic review. *LACLIL*, *11*(1), 19–37.

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.

Green, D. (1999). Mental control in the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*, 67–81.

Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development*, *28*(1), 41–50.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (eds). *Syntax and Sematnics, Vol 3, Speech Acts*. 1975 41–58.

*Grugurović, M., Chapelle, C. A. & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*(2), 165-198.

†Gunzenhauser, C., Karbach, J. & Saalbach, H. (2019). Function of verbal strategies in monolingual vs. bilingual students' planning performance: An experimental approach. *Cognitive Development*, *50*, 1–12.

†Gürkan, S. (2019). Effect of Annotation Preferences of the EFL Students' on Their Level of Vocabulary Recall and Retention. *Journal of Educational Computing Research, 57*(6), 1436-1467.

†Gutiérrez Martínez, A., Ruiz de Zarobe, Y. (2017). Comparing the benefits of a metacognitive reading strategy instruction programme between CLIL and EFL primary school students. *Estudios de lingüística inglesa aplicada*, *17*, 71-92.

Hanan, R. E. (2015). *The effectiveness of explicit grammar instruction for the young foreign language learner: A classroom-based experimental study* (Unpublished doctoral dissertation). University of York, UK.

Hangeun, L., & Hee Kim, K. (2011). Can speaking more languages enhance your creativity? Relationship between bilingualism and creative potential among Korean American students with multicultural link. *Personality and Individual Differences*, *50*, 1186–1190.

Harris, V. (2007). Exploring Progression: Reading and Listening Strategy Instruction with Near-Beginner Learners of French. *Language Learning Journal, 35*(2), 189-204.

**Harris & Ó Duibhir (2011). *Effective language teaching: A synthesis of research*. Report commissioned by the National Council for Curriculum and Assessment. Retrieved 28 Sep 2019 from: https://www.ncca.ie/en/resources/effective_language_teaching_a_synthesis_of_research

*Hassan, X., Macaro, E., Mason, D., Nye, G., Smith, P. & Vanderplank, R. (2005). *Strategy training in language learning – A systematic review of available research*. In: Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Heining-Boynton, A. L., & Haitema, T. (2007). A ten-year chronicle of student attitudes toward foreign language in the elementary school. *Modern Language Journal, 91*, 149–168.

†Heller, M. C., Lervåg, A. & Grøver, V. (2019). Oral language intervention in Norwegian schools serving young language-minority learners: A randomized trial. *Reading Research Quarterly*, *54*(4), 531–552.

†Hennebry, M. & Gao, X. (2018). Interactions between medium of instruction and language learning motivation, *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2018.153019

Hennebry, M., Rogers, V., Macaro, E., & Murphy, V. (2017). Direct teaching of vocabulary after listening: is it worth the effort and what method is best? *The Language Learning Journal, 45*(3), 282-300.

Hermanto, N., Moreno, S., & Bialystok, E. (2012). Linguistic and metalinguistic outcomes of intense immersion education: How bilingual? *International Journal of Bilingual Education and Bilingualism*, *15*(2), 131–145.

Ho, P. V. P., & Binh, N. (2014). The effects of communicative grammar teaching on students' achievement of grammatical knowledge and oral production. *English Language Teaching, 7*(6), 74-86.

Hood, P. (2006). Can early foreign language learning contribute to the shared emotional and motivational landscape of a primary school? *Pastoral Care in Education*, *24*, 4–12.

Hoyt, K. (2016). Developing and evaluating language learners' intercultural competence: Cultivating perspectivetaking. *Dimension*, 2016, 75–102.

*Huang, B. H. (2016). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism*, *13*(3), 257–273.

Huang, K. -J. (2018). On bilinguals' development of metalinguistic awareness and its transfer to L3 learning: The role of language characteristics. *International Journal of Bilingualism*, *22*(3), 330–349.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal, 96*(4), 544-557.

Ibrahim, R., Eviatar, Z., & Aharon-Peretz, J. (2007). Metalinguistic awareness and reading performance: A cross language comparison. *Journal of Psycholinguistic Research*, *36*, 297–317.

Ihle, A., Oris, M., Fagot, D., & Kliegel, M. (2016). The relation of the number of languages spoken to performance in different cognitive abilities in old age. *Journal of Clinical and Experimental Neuropsychology*, *38*(10), 1103–1114.

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.

†Janzen Ulbricht, N. (2018). An experiment on gesture and fluency in two German schools. *ELT Journal*, *72*(3), 309–318.

Jasinska, K. K., & Petitto, L. A. (2013). How age of bilingual exposure can change the neural systems for language in the developing brain: A functional near infrared spectroscopy investigation of syntactic processing in monolingual and bilingual children. *Developmental Cognitive Neuroscience*, *6*, 87–101.

*Jeon, E.-Y. & Day, R. R. (2016). The effectiveness of ER on reading proficiency: A meta-analysis. *Reading in a Foreign Language*, *28*(2), 246–265.

*Jeon, E. H. & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J.M.Norris & L. Ortega (eds). *Synthesizing Research on Language Learning and Teaching*, 165–211. Amsterdam: John Benjamins.

Jiang, Y., & Wang, J. (2018). A study of cultural empathy in foreign language teaching from the perspective of cross-cultural communication. *Theory and Practice in Language Studies*, *8*(12), 1664–1670.

Jiao, L., Lui, C., Wang, R., & Chen, B. (2019). Working memory demand of a task modulates bilingual advantage in executive functions*. International Journal of Bilingualism*, *23*(1), 102–117.

†Kalia, V., Daneri, M. P. & Wilbourn, M. P. (2019). Relations between vocabulary and executive functions in Spanish–English dual language learners. *Bilingualism: Language and Cognition*, *22*(1), 1–14.

*Kang, E. Y., Sok, S., & Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research*, *23*(4), 428–453.

†Karimi, P., Lotfi, A. R. & Biria, R. (2019). Enhancing pilot's aviation English learning, attitude and motivation through the application of Content and Language Integrated learning. *International Journal of Instruction*, *12*(1), 751–766.

†Kasprowicz, R. & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*, *39*(6), 886–911.

†Kasprowicz, R. E., Marsden, E. J., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal, 103*(3), 580-606.

Kaushanskaya, M., Yoo, J., & Marian, V. (2011). The effect of second language experience on native language processing. *Vial–Vigo International Journal of Applied Linguistics*, *8*, 55–77.

Kazemeini, T., & Fadardi, J. S. (2016). Executive function: Comparing bilingual and monolingual Iranian university students*. Journal of Psycholinguistic Research*, *45*(6), 1315–1326.

Kharkhurin, A. V. (2009). The role of bilingualism in creative performance on divergent thinking and invented alien creatures tests*. Journal of Creative Behavior*, *43*, 59–71.

Kharkhurin, A. V. (2010). Bilingual verbal and nonverbal creative behavior. *International Journal of Bilingualism*, *14*, 211–226.

*Kim, J.-K. & Kim, J.-R. (2017). A meta-analysis of content and language integrated learning in English. *Advances in Science, Technology and Engineering Systems Journal*, *2*(3), 1358–1362.

†Košak-Babuder, M., Kormos, J., Ratajczak, M., & Pižorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing*, *36*(1), 51–75.

Kostandyan, M E., & Ledovaya, Y. A. (2013). How the age of language acquisition relates to creativity. *Procedia-Social and Behavioral Sciences*, *86*, 140–145.

Kousaie, S., & Phillips, N. A. (2012). Ageing and bilingualism: Absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. *The Quarterly Journal of Experimental Psychology*, *65*(2), 356–369.

Kroll, J.F. & Dussias, P.E. (2017). The benefits of multilingualism to the personal and professional development of residents of the US. *Foreign Language Annals, 50*, 248-259.

Kuhl, P. K., Stevenson, J., Corrigan, N. M., van den Bosch, J. J. F., Can, D. D., & Richards, T. (2016). Neuroimaging of the bilingual brain: Structural brain correlates of listening and speaking in a second language. *Brain and Language*, *162*, 1–9.

Kuipers, J. -R., & Thierry, G. (2013). ERP-pupil size correlations reveal how bilingualism enhances cognitive flexibility. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *49*(10), 2853–2860.

Kuo, L. J., & Anderson, R. C. (2010). Beyond cross-language transfer: Reconceptualizing the impact of early bilingualism on phonological awareness*. Scientific Studies of Reading*, *14*, 365–385.

Kushalnagar, P., Hannay, H. J., & Hernández, A. E. (2010). Bilingualism and attention: A study of balanced and unbalanced bilingual deaf users of American Sign Language and English. *Journal of Deaf Studies & Deaf Education*, *15*, 263–273.

Kuska, S. K., Zaunbauer, A. C. M., & Möller, J. (2010). Immersion programs in German elementary schools. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *42*, 143–153.

†Kuzminska, N., Stavytska, I., Lukianenko, V. & Lygina, O. (2019). Application of CLIL methodology in teaching economic disciplines at university. *Advanced Education, Special Issue 11*, 112–117.

Ladas, A. I., Carroll, D. J., & Vivas, A. B. (2015). Attentional processes in low-socioeconomic status bilingual children: Are they modulated by the amount of bilingual experience? *Child Development*, *86*(2), 557–578.

Lambert, W.E. & Tucker, G.R. (1972). *The bilingual education of children: The St. Lambert Experiment*. Rowley, MA: Newbury House.

Lan, Y. J., Sung, Y. T., & Chang, K. E. (2009). Let us read together: Development and evaluation of a computer-assisted reciprocal early English reading system. *Computers & Education, 53(*4), 1188-1198.

†Lancaster, N. K. (2018). Extramural exposure and language attainment: The examination of input-related variables in CLIL programmes. *Porta Linguarum*, *29*, 91–114.

†Lau, W. W. F. & Yuen, A. H. K. (2011). The impact of medium of instruction: The case of teaching and learning of computer programming. *Education and Information Technologies*, *16*, 183–201.

Laufer, B. (2006). Comparing Focus on Form and Focus on FormS in second-language vocabulary learning. *Canadian Modern Language Review, 63*(1), 149-166.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716.

Laurent, A., & Martinot, C. (2010). Bilingualism and phonological awareness: The case of bilingual (French-Occitan) children. *Reading and Writing*, *23*, 435–452.

Lazaruk, W. (2007). Linguistic, academic, and cognitive benefits of French immersion. *Canadian Modern Language Review*, *63*, 605–627.

*Lee, S.-K. & Huang, H.-T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, *30*(3), 307–331.

*Lee, J., Jang, J. & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366.

Lee, H., & Kim, K. H. (2010). Relationships between bilingualism and adaptive creative style, innovative creative style, and creative strengths among Korean American students. *Creativity Research Journal*, *22*, 402–407.

Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *Modern Language Journal, 97*(4), 887-901.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, 1–33. http://doi.org/10.1037/bul0000142.

*Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365.

Lichtman, K. (2013). Developmental comparisons of implicit and explicit language learning. *Language Acquisition: A Journal of Developmental Linguistics, 20*(2), 93-108.

Lightbown, P. M. (2008). Easy As Pie? Children Learning Languages. *Concordia Working Papers in Applied Linguistics*, *1*, 5–29.

*Lin, W.-C., Huang, H.-T. & Liou, H.-C. (2013). The effects of text-based SCMC on SLA: A meta analysis. *Language Learning & Technology*, *17*(2), 123–142.

Liu, Y. -C., Liu, Y. -Y., Yip, P. -K., Meguro, M., & Meguro, K. (2017). Speaking one more language in early life has only minor effects on cognition in Taiwanese with low education level: The Taishan project. *Psychogeriatrics*, *17*(4), 256–261.

**Lo, Y. Y. & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, *84*(1), 47–73.

†Lo, Y. Y. & Murphy, V. A. (2010). Vocabulary knowledge and growth in immersion and regular language-learning programmes in Hong Kong. *Language and Education*, *24*(3), 215–238.

Lopes, A. & Cecilia, R.R. (2019). *New trends in foreign language teaching: Methods, evaluation, and innovation.* Newcastle: Cambridge Scholars Publishing.

†Lorge, I. & Katsos, N. (2018) Listener-adapted speech: bilinguals adapt in a more sensitive way. *Linguistic Approaches to Bilingualism*, *9*(3), 376 – 397.

†Lorge, I. & Katsos, N. (2019). Bilinguals adapt in a more sensitive way. *Linguistic Approaches to Bilingualism*, *9*(3), 376–397.

Luan, N. L., & Sappathy, S. M. (2011). L2 vocabulary acquisition: The impact of negotiated interaction. *GEMA Online Journal of Language Studies, 11*(2), 5-20.

*Lyster, R. & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition*, *32*, 265–302.

Macaro, E. (2018). *English Medium Instruction: Content and language in policy and practice*. Oxford: Oxford University Press

Macaro, E. (2003). *Teaching and learning a second language: A guide to current research and its applications*. London: Continuum.

Macaro, E. (2008). The decline in language learning in England: getting the facts right and getting real, *Language Learning Journal*, *36*(1): 101-108. DOI: 10.1080/09571730801988595.

Macaro, E. (2001). *Learning strategies in second and foreign language classrooms*. Continuum: London.

*Macaro, E., Curle, S., Pun, J. & An, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching*, *51*(1), 36–76.

Macaro, E., & Erler, L. (2008). Raising the achievement of young-beginner readers of French through strategy instruction. *Applied Linguistics, 29*(1), 90-119.

Macaro, E., Graham, S. & Woore, R. (2015). *Improving foreign language teaching: Towards a research-based curriculum and pedagogy*. London: Routledge.

*Macaro, E., Handley, Z. & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching*, *45*(1), 1–43.

Macaro, E., & Mutton, T. (2009). Developing reading achievement in primary learners of French: Inferencing strategies versus exposure to 'graded readers'. *Language Learning Journal, 37*(2), 165-182.

Macnamara, B. N., & Conway, A. R. A. (2014). Novel evidence in support of the bilingual advantage: Influences of task demands and experience on cognitive control and working memory. *Psychonomic Bulletin & Review*, *21*(2), 520–525.

*Mahmud, M. M. (2018). Technology and language – What works and what does not: A meta-analysis of blended learning research. *The Journal of AsiaTEFL, 15*(2), 257–565.

Makumane, M. A., & Ngcobo, S. (2018). The socio-economic value of French language education in Lesotho: The learners' voices. *South African Journal of African Languages*, *38*(2), 167–175.

Manoli, P., Papadopoulou, M., & Metallidou, P. (2016). Investigating the immediate and delayed effects of multiple-reading strategy instruction in primary EFL classrooms. *System, 56*, 54-65.

†Marini, A., Eliseeva, N. & Fabbro, F. (2019). Impact of early second-language acquisition on the development of first language and verbal short-term and working memory. *International Journal of Bilingual Education and Bilingualism*, *22*(2), 165–176.

Marinova-Todd, S. H. (2012). "Corplum is a core from a plum": The advantage of bilingual children in the analysis of word meaning from verbal context. *Bilingualism: Language and Cognition*, *15*(1), 117–127.

Marsh, H. W., Hau, K. T., & Kong, C. K. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal*, *39*, 727–763.

Martínez, A. G., & Ruiz de Zarobe, Y. (2017). Comparing the benefits of a metacognitive reading strategy instruction programme between CLIL and EFL primary school students. *Elia, 17*, 71-92.

Marzecová, A., Bukowski, M., Correa, Á., Boros, M., Lupiáñez, J., & Wodniecka, Z. (2013). Tracing the bilingual advantage in cognitive control: The role of flexibility in temporal preparation and category switching. *Journal of Cognitive Psychology*, *25*(5), 586–604.

Mavilidi, M. F., Okely, A. D., Chandler, P., Cliff, D. P., & Paas, F. (2015). Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. *Educational Psychology Review, 27*(3), 413-426.

†Merino, J. A. & Lasagabaster, D. (2018). The effect of content and language integrated learning programmes' intensity on English proficiency: A longitudinal study. *International Journal of Applied Linguistics*, *28*(1), 18–30.

Merisuo-Storm, T. (2007). Pupils' attitudes towards foreign-language learning and the development of literacy skills in bilingual education. *Teaching and Teacher Education*, *23*, 226–235.

†Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B. & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, *39*, 161–188.

†Meyerhöffer, N. & Dreesmann, D. C. (2019a). The exclusive language of science? Comparing knowledge gains and motivation in English-bilingual biology lessons between non-selected and preselected classes. *International Journal of Science Education*, *41*(1), 1–20.

†Meyerhöffer, N. & Dreesmann, D. C. (2019b). English-bilingual biology for standard classes development, implementation and evaluation of an English-bilingual teaching unit in standard German high school classes*. International Journal of Science Education*, *41*(10), 1366–1386.

Mikolic, V. (2010). The relationship between communicative competence and language use in a multicultural environment: The case of Slovene Istria. *Journal of Multilingual and Multicultural Development*, 31, 37–53.

Millar, S. (2017). The sociolinguistic economy in contexts of transience and change in Danish multinational companies. *Journal of Linguistic Anthropology*, 27(3), 344–360.

*Miller, P. C. (2003). *The effectiveness of corrective feedback: A meta-analysis* (Ph.D. dissertation). Purdue University, West Lafayette, Indiana, USA.

Mistar, J., Zuhairi, A., & Yanti, N. (2016). Strategies training in the teaching of reading comprehension for EFL learners in Indonesia. *English Language Teaching, 9*(2), 49- 56.

Mitchell, R., & Myles, F. (1998). *Second language learning theories*. London: Arnold.

Mitchell, R., & Myles, F. (2019). Learning French in the UK setting: Policy, classroom engagement and attainable learning outcomes. *Apples: Journal of Applied Language Studies, 13*(1), 69-93.

Modirkhamene, S. (2006). The reading achievement of third language versus second language learners of English in relation to the interdependence hypothesis. *International Journal of Multilingualism*, *3*, 280–295.

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: a meta-analysis. *System, 41*(3), 720–739.

Mutton, T. & Woore, R. (2014). Designing tasks to promote learning in the foreign language classroom, In: I Thompson (ed.) *Designing tasks in secondary education: Enhancing subject understanding and student engagement*. Abingdon: Routledge.

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

Murphy, V.A. (2018). Literacy development in linguistically diverse pupils. In D. Miller, F. Bayram, J. Rothman & L. Serratrice (Eds). Bilingual Cognition and Language: The state of the science across its subfields. *Studies in Bilingualism, 54*, Amsterdam: John Benjamins.

Murphy, V.A. (2019). Multilingualism in primary schools. In S. Garton & F. Copland (Eds). *The Routledge Handbook of Teaching English to Young Learners*. Abingdon: Taylor & Francis.

Murphy, V.A. & Evangelou, M. (Eds). (2015). *Early childhood education in English for speakers of other languages*. London: British Council.

Murphy, V.A. & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. London: Education Endowment Foundation

Murphy, V.A., Macaro, E., Cipolla, C. & Alba, S. (2015). The influence of L2 learning on first language literacy skills. Applied Psycholinguistics, *36*(5), 1133-1153 doi:10.1017/S0142716414000095

†Nemati, F., Ghaemi, F., Amini, M. & Mohamadi, Z. (2017). The impact of English versus Persian songs on Iranian EFL learners' mastery of English letters. *International Journal of Language Studies*, *11*(2), 67–88.

Netten, J., & Germain, C. (2009). The future of intensive French in Canada. *The Canadian Modern Language Review*, *65*(5), 757-786.

Ngo, C. M., & Trinh, L. Q. (2011). Lagging behind writing pedagogical developments: the impact of implementing process-based approach on learners' writing in a Vietnamese secondary education context. *Journal on English Language Teaching, 1*(3), 60-71.

*Norris, J. M. & Ortega, L. (2008). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528.

*Norris, J. M. & Ortega, L. (2005). Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning*, *51*(s1), 157–213.

Ong, G., Sewell, D. K., Weekes, B., McKague, M., & Abutalebi, J. (2017). A diffusion model approach to analysing the bilingual advantage for the Flanker task: The role of attentional control processes. *Journal of Neurolinguistics*, 43, 28–38.

Onnis, L., Chun, W. E., & Lou-Magnuson, M. (2018). Improved statistical learning abilities in adult bilinguals. *Bilingualism*, *21*(2), 427–433.

Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagardmid, A. (2016). Rayyan – a web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 210.

†Owen, E. A., Razali, A. B., Abd Samad, A., & Noordin, N. (2019). Enhancing Libyan students' English speaking performance through language game and information gap activities. *Problems of Education in the 21st Century, 77*(1), 110-125.

Oxley, E. & De Cat, C. (2019). A systematic review of language and literacy interventions in children and adolescents with English as an Additional Language. *The Language Learning Journal*, https://doi.org/10.1080/09571736.2019.1597146

†Paap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S., & Mason, L. (2018) On the encapsulation of bilingual language control. *Journal of Memory and Language, 105*, 76–92.

†Padial-Ruz, R., García-Molina, R., & Puga-González, E. (2019). Effectiveness of a Motor Intervention Program on Motivation and Learning of English Vocabulary in Preschoolers: A Pilot Study. *Behavioural Sciences, 9*(8), 84.

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals*, *46*, 661–679.

†Papageorgiou, A., Bright, P., Periche Tomas, E., & Filippi, R. (2019). Evidence against a cognitive advantage in the older bilingual population. *Quarterly Journal of Experimental Psychology*, *72*(6), 1354–1363.

†Paplikar, A., Mekala, S., Bak, T. H., Dharamkar, S., Alladi, S. & Kaul, S. (2019). Bilingualism and the severity of poststroke aphasia. *Aphasiology*, 33(1), 58–72.

Park, J., Ellis Weismer, S., & Kaushanskaya, M. (2018). Changes in executive function over time in bilingual and monolingual school-aged children. *Developmental Psychology*, *54*(10), 1842–1853.

†Park, A. Y., Isaacs, T. & Woodfield, H. (2018). A comparison of the effects of extensive and intensive reading approaches on the vocabulary development of Korean secondary EFL learners. *Applied Linguistics Review*, *9*(1), 113–134.

Perani, D., Farsad, M., Ballarini, T., Lubian, F., Malpetti, M., Fracchetti, A., & Abutalebi, J. (2017). The impact of bilingualism on brain reserve and metabolic connectivity in Alzheimers' dementia. *Proceedings of the National Academy of Sciences*, *114*, 1690–1695.

*Perez, M. M. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739.

Perquin, M., Diederich, N., Pastore, J., Lair, M. -L., Stranges, S., & Vaillant, M. (2015). Prevalence of dementia and cognitive complaints in the context of high cognitive reserve: A population-based study. *PLOS One*, *10*(9).

*Persson, V. & Nouri, J. (2018). A systematic review of second language learning with mobile technologies. *International Journal of Emerging Technologies in Learning*, *13*(2), 188–210.

†Pfenninger, S. E. & Singleton, D. (2019). Starting age overshadowed: The primacy of differential environmental and family support effects on second language attainment in an instructional context*.* *Language Learning*, *69*(S1), 207–234.

†Pladevall-Ballester, E. (2019). A longitudinal study of primary school EFL learning motivation in CLIL and non-CLIL settings. *Language Teaching Research*, *23*(6), 765–786.

*Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*(4), 993–1038.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. & Duffy, S. (2006). Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. Retrieved 13 Jan 2010 from: https://www.researchgate.net/publication/233866356_Guidance_on_the_conduct_of_narrative_synthesis_in_systematic_reviews_A_product_from_the_ESRC_Methods_Programme

Porter, A. (2016). A helping hand with language learning: teaching French vocabulary with gesture. *Language Learning Journal, 44*(2), 236-256.

Poulin-Dubois, D., Blaye, A., Coutya, J., & Bialystok, E. (2011). The effects of bilingualism on toddlers' executive functioning. *Journal of Experimental Child Psychology*, *108*, 567–579.

Prior, A., & MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism and Cognition*, *13*, 253–262.

†Pujadas, G. & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, *47*(4), 479–496.

†Raap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S. & Mason, L. (2019). On the encapsulation of bilingual language control*. Journal of Memory and Language*, *105*, 76–92.

Ransdell, S., Barbier, M., & Niit, T. (2006). Metacognitions about language skill and working memory among monolingual and bilingual college students: When does multilingualism matter? *International Journal of Bilingual Education & Bilingualism*, *9*, 728–741.

Reder, F., Marec-Breton, N., Gombert, J. -E., & Demont, E. (2013). Second-language learners' advantage in metalinguistic awareness: A question of languages' characteristics. *British Journal of Educational Psychology*, *83*(4), 686–702.

Richards, J. & Rogers, T.S. (2014). *Approaches and Methods in Language teaching, 3rd edition*. Cambridge: Cambridge University Press.

†Robinson, M. G. & Sorace, A. (2019). The influence of collaborative language learning on cognitive control in unbalanced multilingual migrant children. *European Journal of Psychology of Education*, *24*, 255–272.

†Rostamian, M., Fazilatfar, A. M., & Jabbari, A. A. (2018). The effect of planning time on cognitive processes, monitoring behavior, and quality of L2 writing. *Language Teaching Research, 22(*4), 418–438.

*Rubio-Alcalá, F. D., Arco-Tirado, J. L., Fernández-Martín, F. D., López-Lechuga, R., Barrios, E. & Pavón-Vázquez, V. (2019). A systematic review on evidences supporting quality indicators of bilingual, plurilingual and multilingual programs in higher education. *Educational Research Review*, *27*, 191–204.

†Ruiz de Zarobe, Y. & Zenotz, V. (2015). Reading strategies and CLIL: The effect of training in formal instruction. *Language Learning Journal*, *43*(3), 319–333.

†Ruiz de Zarobe, Y. & Zenotz, V. (2018). Learning strategies in CLIL classrooms: How does strategy instruction affect reading competence over time*? International Journal of Bilingual Education and Bilingualism*, *21*(3), 319–331.

†Safataj, M. & Amiryousefi, M. (2016). Effect of homonymous set of words instruction on vocabulary development and retention of young female elementary learners in Iranian EFL context through metalinguistic awareness. *Theory and Practice in Language Studies*, *6*(11), 2092–2101.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. TESOL Quarterly*, *46*(4), 842–854.

*Saito, K. & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708.

Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *The Review of Economics and Statistics*, *87*, 523–538.

†Salvador-García, C., Chiva-Bartoll, O. & Capella-Peris, C. (2019). Bilingual physical education: The effects of CLIL on physical activity levels. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2019.1639131

Salvatierra, J. L., & Rosselli, M. (2011). The effect of bilingualism and age on inhibitory control. *International Journal of Bilingualism*, *15*, 26–37.

†San Isidro, X., & Lasagabaster, D. (2019). The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research, 23*(5), 584–602.

Santillán, J., & Khurana, A. (2018). Developmental associations between bilingual experience and inhibitory control trajectories in Head Start children. *Developmental Science, 21*(4), 1–12.

Schmidt, R. W. (1990) 'The Role of Consciousness in Second Language Learning'. *Applied Linguistics*, *11*(2): 129–158. DOI: https://doi.org/10.1093/applin/11.2.129

Schoenpflug, U., & Klische, F. (2010). Cross- and mono-linguistic text processing in bilingual children. *Educational Psychology, 30*, 849–870.

Schroeder, S. R., & Marian, V. (2012). A bilingual advantage for episodic memory in older adults. *Journal of Cognitive Psychology, 24*(5), 591–601.

Sercu, L. (2013). Weblogs in foreign language education: real and promised benefits. In L. G. Chova, A. L. Martínez, & I. C. Torres (Eds.), *7th International Technology, Education and Development Conference* (pp. 4355-4366).

†Serratrice L, De Cat C (2019). Individual differences in the production of referential expressions: The effect of language proficiency, language exposure and executive function in bilingual and monolingual children. *Bilingualism: Language and Cognition*, 1–16.

*Sharifi, M. (2018). Retrospect and prospect of computer assisted English language learning: A meta-analysis of the empirical literature. *Computer Assisted Language Learning, 31*(4), 413–436.

†Shi, T. (2018). A study of the TPR method in the teaching of English to primary school students. *Theory and Practice in Language Studies, 8*(8), 1087–1093.

†Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research, 15*(2), 137–158.

Shintani, N. (2012). Repeating input-based tasks with young beginner learners. *RELC Journal, 43*(1), 39-51.

Shintani, N. (2013). The effect of focus on form and focus on forms instruction on the acquisition of productive knowledge of L2 vocabulary by young beginning-level learners. *TESOL Quarterly, 47*(1), 36-62.

*Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, *36*(3), 306–325.

Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., & Itakura, S. (2010). Bilingualism accentuates children's conversational understanding. *PLoS One*, 5, e9004.

Silven, M., & Rubinov, E. (2010). Language and preliteracy skills in bilinguals and monolinguals at preschool age: Effects of exposure to richly inflected speech from birth. *Reading and Writing*, *23*, 385–414.

†Singh, L., Quinn, P. C., Xiao, N. G. & Lee, K. (2019). Monolingual but not bilingual infants demonstrate racial bias in social cue use. *Developmental Science*, *22*(6), e12809.

Soveri, A., Laine, M., Hamalainen, H., & Hugdahl, K. (2011). Bilingual advantage in attentional control: Evidence from the forced-attention dichotic listening paradigm. *Bilingualism, Language and Cognition*, *14*, 371–378.

*Spada, N. & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 263–308.

Stafford, C. A. (2011). Bilingualism and enhanced attention in early adulthood. *International Journal of Bilingual Education & Bilingualism*, *14*, 1–22.

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of Dual-Language Immersion Programs on Student Achievement. *American Educational Research Journal*, *54*(1_suppl), 282S–306S. http://doi.org/10.3102/0002831216634463

Stephens, S., & Moxham, B. J. (2019). Do medical students who are multilingual have higher spatial and verbal intelligence and do they perform better in anatomy examinations? Multilingualism and spatial and verbal intelligence. *Clinical Anatomy*, *32*(1), 26–34.

Struys, E., Duyck, W., & Woumans, E. (2018). The role of cognitive development and strategic task tendencies in the bilingual advantage controversy. *Frontiers in Psychology*, *9*(1790), 1–11.

†Suárez, M. M. & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference*? The Language Learning Journal*, *47*(4), 497–517.

Sun, B., Hu, G., & Curdt-Christiansen, X. L. (2018). Metalinguistic contribution to writing competence: A study of monolingual children in China and bilingual children in Singapore. *Reading and Writing: An Interdisciplinary Journal*, *31*(7), 1499–1523.

Sun, X., Li, L., Ding, G., Wang, R., & Li, P. (2019). Effects of language proficiency on cognitive control: Evidence from resting-state functional connectivity. *Neuropsychologia*, *129*, 263–275.

*Sung, Y.-T., Chang, K.-E. & Yang, J.-M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, *16*, 68–84.

*Taj, I. H., Sulan, N. B., Sipra, M. A. & Ahmad, W. (2016). Impact of mobile assisted language learning (MALL) on EFL: A meta-analysis. *Advances in language and literary studies*, *7*(2), 76–83.

Takeda, C. (2002). The application of phonics to the teaching of reading in junior high school English classes in Japan. *TESL Reporter, 35*(2), 16-36.

Tammenga-Helmantel, M., Arends, E., & Canrinus, E. T. (2014). The effectiveness of deductive, inductive, implicit and incidental grammatical instruction in second language classrooms. *System, 45*, 198-210.

Taylor, C., & Lafayette, R. (2010). Academic achievement through FLES: A case for promoting greater access to foreign language study among young learners. *Modern Language Journal*, *94*, 22–42.

Taylor, A., Lazarus, E., & Cole, R. (2005). Putting languages on the (drop down) menu: innovative writing frames in modern foreign language teaching. *Educational Review, 57*(4), 435-455.

*Taylor, A., Stevens, J. R. & Asher, J. W. (2006). The effects of explicit reading strategy training on L2 reading comprehension. In J.M. Norris & L. Ortega (eds). *Synthesizing Research on Language Learning and Teaching*, 213–244. Amsterdam: John Benjamins.

†Teng, F. (2019a). A comparison of text structure and self-regulated strategy instruction for elementary school students' writing. *English Teaching: Practice & Critique*, *18*(3), 281–297.

†Teng, F. (2019b). Incidental vocabulary learning for primary school students: the effects of L2 caption type and word exposure frequency. *The Australian Educational Researcher*, *46*, 113–136.

†Teng, F. (2019c). Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning*, *32*(7), 665–691.

Thompson, A. S. (2013). The interface of language aptitude and multilingualism: Reconsidering the bilingual/multilingual dichotomy. *Modern Language Journal*, *97*(3), 685–701.

Tode, T. (2007). Durability problems with explicit instruction in an EFL context: the learning of the English copula "be" before and after the introduction of the auxiliary "be". *Language Teaching Research, 11*(1), 11-30.

Torres-Zúñiga, L. & Schmidt, T.H. (Eds). (2017). *New methodological approaches to foreign language teaching.* Newcastle: Cambridge Scholars Publishing.

Toth, P. D., & Guijarro-Fuentes, P. (2013). The impact of instruction on second-language implicit knowledge: Evidence against encapsulation. *Applied Psycholinguistics, 34*(6), 1163-1193.

†Tran, C. D., Arredondo, M. M. & Yoshida, H. (2019). Early executive function: The influence of culture and bilingualism. *Bilingualism: Language and Cognition*, *22*(4), 714–732.

*Tsai, Y.-L. & Tsai, C.-C. (2018). Digital game-based second-language vocabulary learning and conditions of research designs: A meta-analysis study. *Computers & Education, 125*, 345–357.

Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning, 27*(1), 1-25.

Valdés, G., Kibler, A. & Philipose, S. (2004). *What does research show about the benefits of language learning?* Retrieved from: https://www.actfl.org/advocacy/what-the-research-shows

*Vahedi, V. S., Ghonsooly, B. & Pishghadam, R. (2016). Vocabulary glossing: A meta-analysis of the relative effectiveness of different gloss types on L2 vocabulary acquisition. *Teaching English with Technology*, *16*(2), 3–25.

†Valis, M., Slaninova, G., Prazak, P., Poulova, P. Kacetl, J. & Klimova, B. (2019). Impact of learning a foreign language on the enhancement of cognitive functions among healthy older population. *Journal of Psycholinguistic Research*, *48*(6), 1211–1318.

†Van de Guchte, M., Rijlaarsdam, G., Braaksma, M., & Bimmel, P. (2019). Focus on language versus content in the pre-task: Effects of guided peer-video model observations on task performance. *Language Teaching Research*, *23*(3), 310–329.

†van de Ven, M., Segers, E. & Verhoeven, L. (2019). Enhanced second language vocabulary learning through phonological specificity training in adolescents. *Language Learning*, *69*(1), 222–250.

†van Veen, S., Remmers, S., Aarnoudse-Moens, C. S. H., Oosterlaan, J., van Kaam, A. H., van Wassenaer-Leemhuis, A. G. (2019). Multilingualism was associated with lower cognitive outcomes in children who were born very and extremely preterm. *Acta Paediatr*, *108*(3), 479–485.

Vanderplank, R. (2010).Déjà Vu? A Decade of Research on Language Laboratories, Television and Video in Language Learning. *Language Teaching 43(1)*, 1-37.

Vanderplank, R. (2012). Plus ça change....Why Latin works for Key Stage 3 pupils and French doesn't. TES, August, 2012. https://www.tes.com/news/modern-foreign-languages-plus-ca-change

*Valezy, J. R. & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J.M. Norris & L. Ortega (eds). *Synthesizing Research on Language Learning and Teaching*, 133–164. Amsterdam: John Benjamins.

Verhagen, J., Grassmann, S., & Küntay, A. C. (2017). Monolingual and bilingual children's resolution of referential conflicts: Effects of bilingualism and relative language proficiency. *Cognitive Development*, *41*, 10–18.

Verhoeven, L. (2007). Early bilingualism, language transfer, and phonological awareness. *Applied Psycholinguistics*, *28*, 425–439.

†Vyn, R., Wesely, P.M. & Neubauer, D. (2019). Exploring the effects of foreign language instructional practices on student proficiency development. *Foreign Language Annals*, *52*(1), 45–65.

†Wang, F., Hwang, W-Y., Li, Y-H., Chen, P-T., & Manabe, K. (2019). Collaborative kinesthetic EFL learning with collaborative total physical response. *Computer Assisted Language Learning, 32*(7) 745-783.

Whiteside, K.E., Gooch, D. & Norbury, C. (2017). English language proficiency and early school attainment among children learning English as an additional language. *Child Development, ii(3)*, 812-827.

Williams, N., & Thomas, E. M. (2017). Exploring minority language input sources as means of supporting the early development of second language vocabulary and grammar. *Applied Psycholinguistics, 38*(4), 855-880.

Winch, C.A., Oancea, A. & Orchard, J. (2015). The contribution of educational research to teachers' professional learning: philosophical understandings. *Oxford Review of Education, 41* (2): 202-216.

Wivers, W.M. (2018). *Teaching foreign language skill, 2nd edition*. Chicago: Chicago University Press.

Wodniecka, Z., Craik, F.I.M., Lin, L. & Bialystok, E. (2010). Does bilingualism help memory? Competing effects of verbal ability and executive control. *International Journal of Bilingual Education & Bilingualism*, *13*, 575–595.

*Woll, B. & Wei, L. (2019). Cognitive benefits of language learning: Broadening our perspectives. Report commissioned by the British Academy. Retrieved 29 Sep 2019 from: https://www.thebritishacademy.ac.uk/sites/default/files/Cognitive-Benefits-Language-Learning-Final-Report.pdf

Woore, R, Graham, S, Porter, A, Courtney, L, Savory, C (2018) *Foreign Language Education: Unlocking Reading (FLEUR) - A study into the teaching of reading to beginner learners of French in secondary school*. Accessible at: https://ora.ox.ac.uk/objects/uuid:4b0cb239-72f0-49e4-8f32-3672625884f0.

Woumans, E., Ameloot, S., Keuleers, E., & Van Assche, E. (2019). The relationship between second language acquisition and nonverbal cognitive abilities. *Journal of Experimental Psychology: General*, *148*(7), 1169–1177.

Woumans, E., Santens, P., Sieben, A., Versijpt, J., Stevens, M., & Duyck, W. (2015). Bilingualism delays clinical manifestation of Alzheimer's disease. *Bilingualism: Language and Cognition, 18*(3), 568–574.

Yang, S., & Yang, H. (2016). Bilingual effects on deployment of the attention system in linguistically and culturally homogeneous children and adults. *Journal of Experimental Child Psychology*, *146*, 121–136.

Yeldham, M. (2018). Viewing L2 captioned videos: What's in it for the listener? *Computer Assisted Language Learning, 31(4)*, 367-389

Yeon, S., Bae, H. S., & Joshi, R. M. (2017). Cross-language transfer of metalinguistic skills: Evidence from spelling English words by Korean students in grades 4, 5 and 6. *Dyslexia: An International Journal of Research and Practice*, *23*(4), 428–448.

*Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*(1), 39–58.

Yunus, M. M., Nordin, N., Salehi, H., Embi, M. A., & Salehi, Z. (2013). The use of information and communication technology (ICT) in teaching ESL writing skills. *English Language Teaching, 6*(7), 1-8.

Zahodne, L. B., Schofield, P. W., Farrell, M. T., Stern, Y., & Manly, J. J. (2014). Bilingualism does not alter cognitive decline or dementia risk among Spanish-speaking immigrants. *Neuropsychology*, *28*(2), 238–246.

Zaunbauer, A. C. M., & Möller, J. (2010). Academic development of children in an immersion program: Results for the first two years of school. *Psychologie In Erziehung Und Unterricht*, *57*, 30–45.

*Zhang, W. & Cheung, Y. L. (2018). Researching innovations in English language writing instruction: A state-of-the-art review. *Journal of Language Teaching and Research*, *9*(1), 80-89.

†Ziegler, N. A. (2014). Fostering self-regulated learning through the European Language Portfolio: An embedded mixed methods study. *The Modern Language Journal*, *98*(4), 921–936.

6. Appendices

6.1 Appendix 1: Email communication with professional networks in language education

The email below was sent to the membership listservs of the following professional networks:

- The Research in Primary Languages (RiPL) Network: www.ripl.uk/network
- The British Association of Applied Linguistics (BAAL): www.baal.org.uk
- The Association Internationale de Linguistique Appliquée (AILA): aila.info
- The National Centre for Excellence for Language Pedagogy (NCELP): ncelp.org

Dear colleagues,

The Applied Linguistics Research Group at the Department of Education at Oxford is currently working on a Rapid Evidence Assessment, funded by the Education Endowment Foundation, of research into languages in education and foreign language teaching and learning.

Our specific review questions are:

- What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
- What is the impact of learning a foreign language on students' wider academic outcomes?
- What practitioner skills or programme characteristics contribute to effective language learning among students?
- What is the impact of using a non-native language as the medium of instruction in academic subjects on students' academic outcomes?
- Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?
- What is the impact of delaying or accelerating the introduction of a new 'local' language as a medium of instruction for new arrivals (e.g., refugees, immigrants) who are not yet proficient in their native language?

To that end I am emailing to ask members of [name of organisation] to email to us details of any research of which they are aware, either published via traditional channels or constituting 'grey literature' (e.g., working papers; PhD theses), that fits the following criteria:

Research which was

- Published / distributed / released in or after 2000

AND which is a

- Systematic review;
- Meta-analysis;
- Narrative review

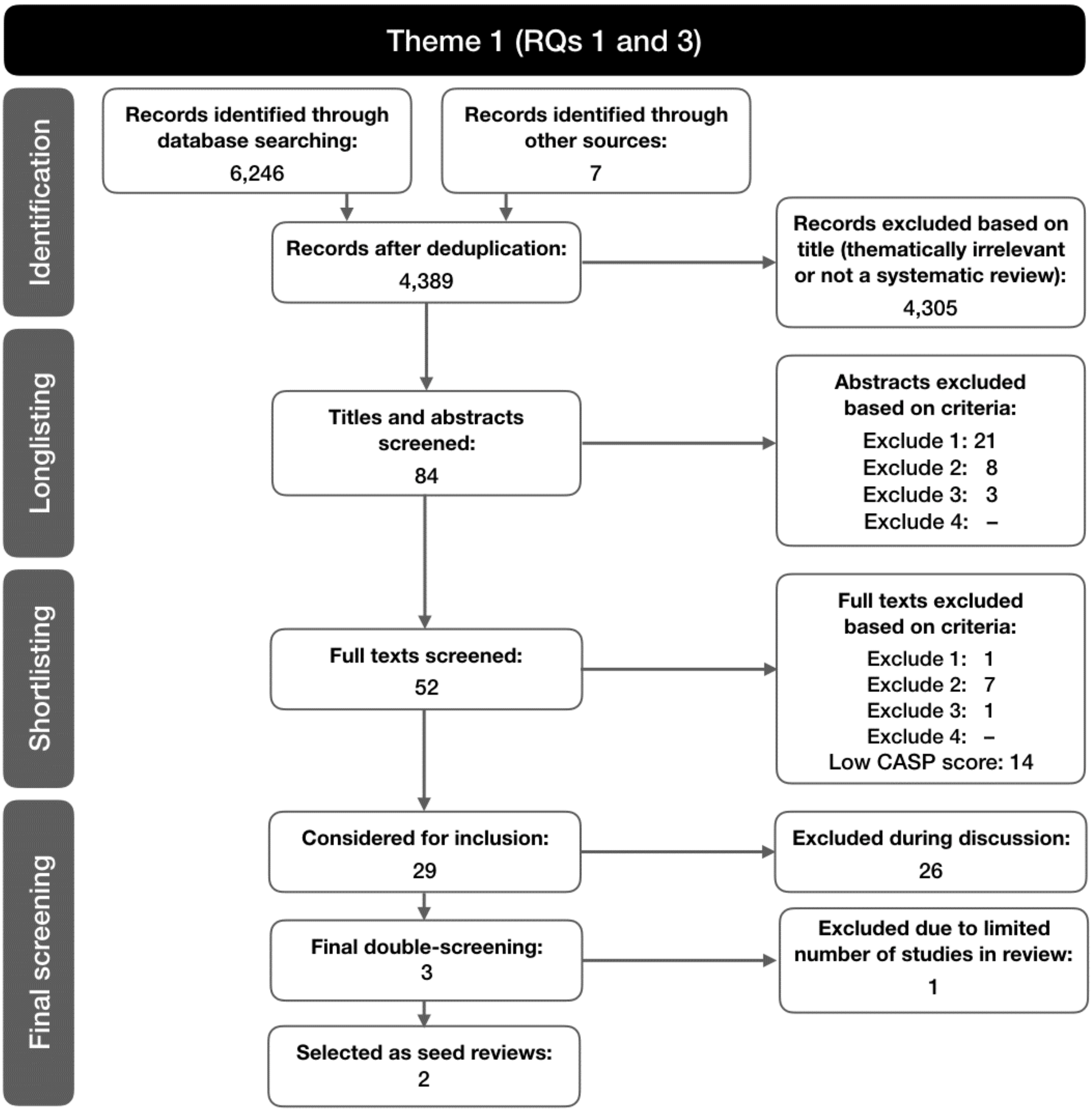
- State-of-the-art article/review; or
- Rapid evidence assessment

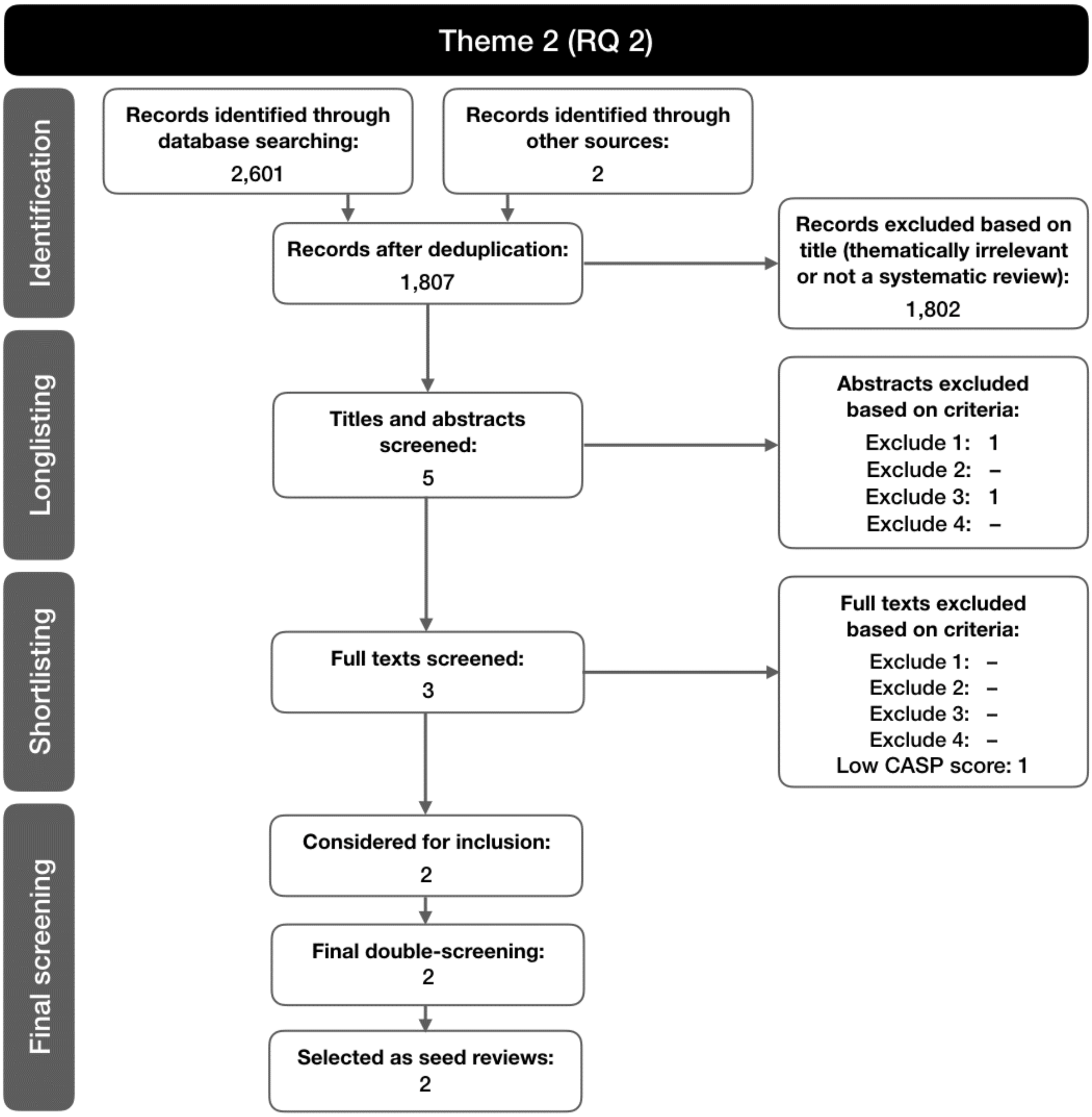
AND which focuses on

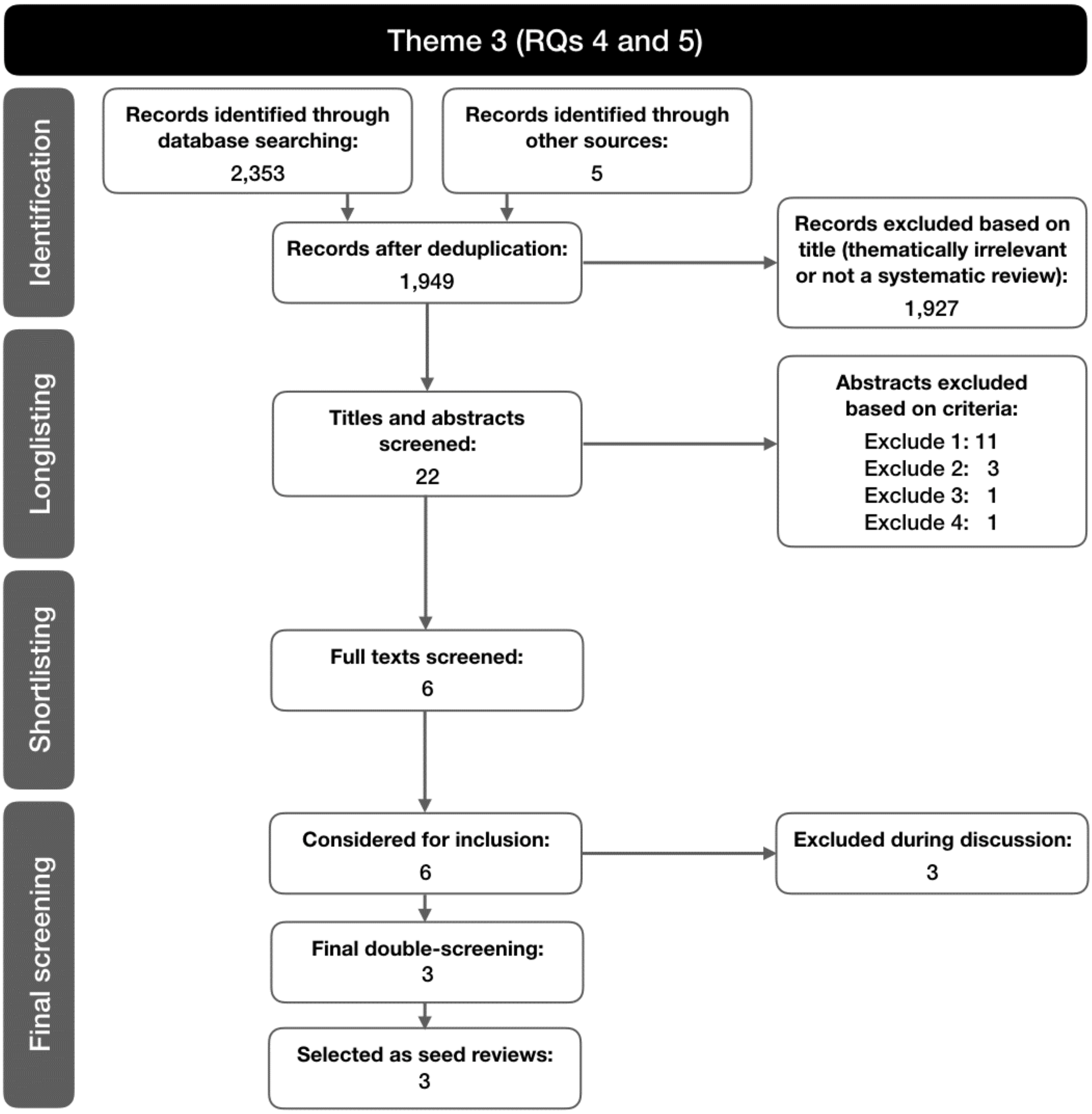
- Language(s) in education;
- Second / foreign / additional language learning; and/or
- Second / foreign / additional language teaching

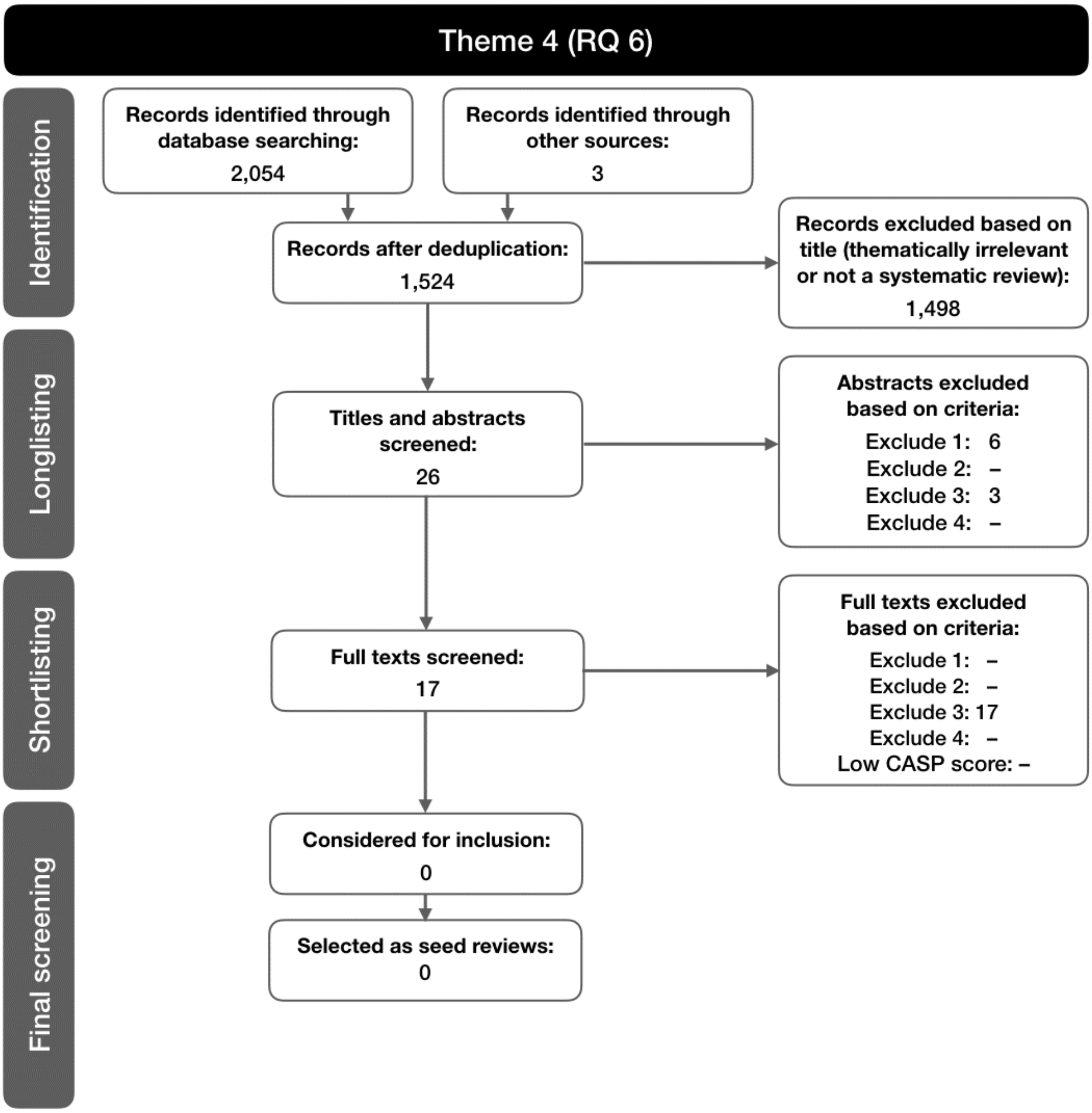
In addition, if you know of any other research (i.e. stand-alone studies) that speaks directly to our review questions, we would be grateful for details of these. Please kindly email references (or weblinks, full text documents etc.) of any relevant studies to Henriette Arndt [email address] at your earliest convenience and thank you for your help!

6.2 Appendix 2: Flow diagrams showing the number of systematic reviews identified and screened per research theme


Theme 1 (RQs 1 and 3)
Identification
Longlisting
Shortlisting
Final screening
Records identified through database searching: 6,246
Records identified through other sources: 7
Records after deduplication: 4,389
Records excluded based on title (thematically irrelevant or not a systematic review): 4,305
Titles and abstracts screened: 84
Abstracts excluded based on criteria:
Exclude 1: 21
Exclude 2: 8
Exclude 3: 3
Exclude 4: –
Full texts screened: 52
Full texts excluded based on criteria:
Exclude 1: 1
Exclude 2: 7
Exclude 3: 1
Exclude 4: –
Low CASP score: 14
Considered for inclusion: 29
Excluded during discussion: 26
Final double-screening: 3
Excluded due to limited number of studies in review: 1
Selected as seed reviews: 2


Theme 2 (RQ 2)
Identification
Longlisting
Shortlisting
Final screening
Records identified through database searching: 2,601
Records identified through other sources: 2
Records after deduplication: 1,807
Records excluded based on title (thematically irrelevant or not a systematic review): 1,802
Titles and abstracts screened: 5
Abstracts excluded based on criteria:
Exclude 1: 1
Exclude 2: –
Exclude 3: 1
Exclude 4: –
Full texts screened: 3
Full texts excluded based on criteria:
Exclude 1: –
Exclude 2: –
Exclude 3: –
Exclude 4: –
Low CASP score: 1
Considered for inclusion: 2
Final double-screening: 2
Selected as seed reviews: 2
]


Theme 3 (RQs 4 and 5)
Identification
Longlisting
Shortlisting
Final screening
Records identified through database searching: 2,353
Records identified through other sources: 5
Records after deduplication: 1,949
Records excluded based on title (thematically irrelevant or not a systematic review): 1,927
Titles and abstracts screened: 22
Abstracts excluded based on criteria:
Exclude 1: 11
Exclude 2: 3
Exclude 3: 1
Exclude 4: 1
Full texts screened: 6
Considered for inclusion: 6
Excluded during discussion: 3
Final double-screening: 3
Selected as seed reviews: 3


Theme 4 (RQ 6)
Identification
Longlisting
Shortlisting
Final screening
Records identified through database searching: 2,054
Records identified through other sources: 3
Records after deduplication: 1,524
Records excluded based on title (thematically irrelevant or not a systematic review): 1,498
Titles and abstracts screened: 26
Abstracts excluded based on criteria:
Exclude 1: 6
Exclude 2: –
Exclude 3: 3
Exclude 4: –
Full texts screened: 17
Full texts excluded based on criteria:
Exclude 1: –
Exclude 2: –
Exclude 3: 17
Exclude 4: –
Low CASP score: –
Considered for inclusion: 0
Selected as seed reviews: 0

6.3 Appendix 3: Terms for electronic database searches

RQ1. What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
RQ3. What practitioner skills or programme characteristics contribute to effective language learning among students?

Type of publication	Systematic review OR meta-analysis OR rapid evidence assessment OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR AND
Context	second language* OR foreign language* OR modern language* OR additional language* OR L2 OR MFL AND teach* OR learn* OR instruct* OR pedagog* OR acqui* OR train* OR study* OR educat* OR intervention* AND
Outcomes	effect* OR outcome* achiev* OR improv* OR develop* OR attain* OR gain* OR increas* OR grow* OR succe* OR competenc* OR develop* OR scor* OR grad* OR result*

RQ2. What is the impact of learning a foreign language on students' wider academic outcomes?

Type of publication	systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment AND
Context	second language* OR foreign language* OR modern language* OR additional language* OR L2 OR MFL AND teach* OR learn* OR instruct* OR pedagog* OR acqui* OR train* OR study* OR educat* AND
Outcomes	effect* OR affect* OR impact* OR influence* OR improv* OR promot* OR benefit* AND academic OR achiev* OR outcome* OR skill* OR literac* OR read* OR competenc* OR metacognit* OR meta-cognit* OR metalinguistic OR meta-linguistic OR problem-solv* OR develop* OR scor* OR grad* OR result* OR subject* OR motivat* OR exam* OR set* OR band* OR GCSE* OR GRE* OR level* OR SAT* OR baccalaureate

RQ4. What is the impact of using a non-native language as the medium of instruction in academic subjects on students' academic outcomes?
RQ5. Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?

Type of publication	systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment

	AND
Context	EMI OR CLIL OR medium instruct* OR medium educat* OR content and language OR content-and-language OR content-based OR bilingual educat* OR bilingual program* OR bilingual programme* OR language immersion OR dual language OR dual-language OR content-based education OR content-based education OR L2MI
	AND
Outcomes	effect* OR outcome* OR achiev* OR improv* OR develop* OR attain* OR gain* OR increas* OR grow* OR succe* OR competenc*

RQ6. What is the impact of delaying or accelerating the introduction of a new 'local' language as a medium of instruction for new arrivals (e.g. refugees, immigrants) who are not yet proficient in their native language?

Type of publication	systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment
	AND
Context	additional language OR EAL OR English language learner* OR new arrival* OR refugee* OR immigrant* OR minority-language* OR language-minorit* OR language minorit* OR linguistic minorit* OR multicultural OR mother tongue* OR native language* OR heritage language*
	AND
Outcomes	effect* OR affect* OR impact* OR influence* OR improv* OR promot* OR benefit* AND academic OR achiev* OR outcome* OR skill* OR literac* OR read* OR competenc* OR metacognit* OR meta-cognit* OR metalinguistic OR meta-linguistic OR problem-solv* OR develop* OR scor* OR grad* OR result* OR subject* OR motivat* OR exam* OR set* OR band* OR GCSE* OR GRE* OR level* OR SAT* OR baccalaureate

6.4 Appendix 4: Adapted CASP checklist for assessing the quality of Systematic Reviews

Criterion	Prompt	Response Scale: Lower	Response scale: Upper
Reference	Include the full citation of the paper being assessed.		
1. Relevance	How relevant are the questions addressed by this review to the objectives of the REA? If the review is not relevant (score = 0), stop here.	0 = Not at all relevant	5 = Very relevant
2. Focus	Did the review address a clearly focused question? An issue can be 'focused' in terms of • the population studied • the intervention given • the outcome considered	0 = Not focused at all	5 = Very clearly focused
3. Sources	Did the authors look for the right type of papers? The 'right type of papers' would • clearly address the review's questions • have an appropriate study design (RCTs or QEDs as per the EEF's guidelines)	0 = No, not at all	5 = Yes, definitely
4. Methods	Are the search methods used to identify the relevant studies described in enough detail to permit replication? Descriptions should include • the search date • the databases used • the search strategy • the search terms	0 = No, not enough detail	5 = Yes, very clearly described
5. Inclusiveness	Do all important, relevant studies seem to be included? Look for • which bibliographic databases were used • follow up from reference lists • personal contact with experts • unpublished as well as published studies	0 = Very limited sources or not clearly described	5 = Exhaustive sources, clearly described

	• non-English language studies		
6. Quality	Did the review's authors do enough to assess the quality of the included studies? The authors need to consider the rigour of the studies they have identified. Did they conduct high standard Risk of Bias or Weight of Evidence assessment procedures?	0 = Substandard assessment or not reported	5 = High quality assessment, clearly reported
7. Synthesis	If the results have been combined, was it reasonable to do so? Consider whether • results of all the included studies are clearly displayed • results were similar from study to study • reasons for variations in results are discussed	0 = No, not appropriate at all	5 = Yes, entirely appropriate
8. Clarity	Are the overall results of the review clear and precise? Consider whether • the 'bottom line results' are clearly presented and in what way • the results are expressed precisely (e.g. numerically, if appropriate) • the stated conclusions are supported by the data presented	0 = No, extremely difficult to understand	5 = Yes, entirely clear and precise
9. Precision	How statistically precise are the results? Where appropriate, look at the confidence intervals (if provided).	0 = No statistical information is provided	5 = Precise statistical information is provided, including confidence intervals where appropriate
10. Applicability	Can the results be applied to the current REA? • Which review question(s) do they address? • How strong is the evidence? • Are there any mediators to consider?	0 = No, the results are not relevant at all	5 = Yes, the results are highly relevant
11. Comprehen-siveness	Were all important outcomes considered? Consider whether there is other information you would have liked to have seen.	0 = No, there are significant omissions	5 = Yes, all relevant information is covered
12. Overall Score	Is this review of high enough quality and sufficient to inform the REA?	0 = Not at all relevant and/or of the lowest quality	5 = Yes, it is exactly relevant and of high quality

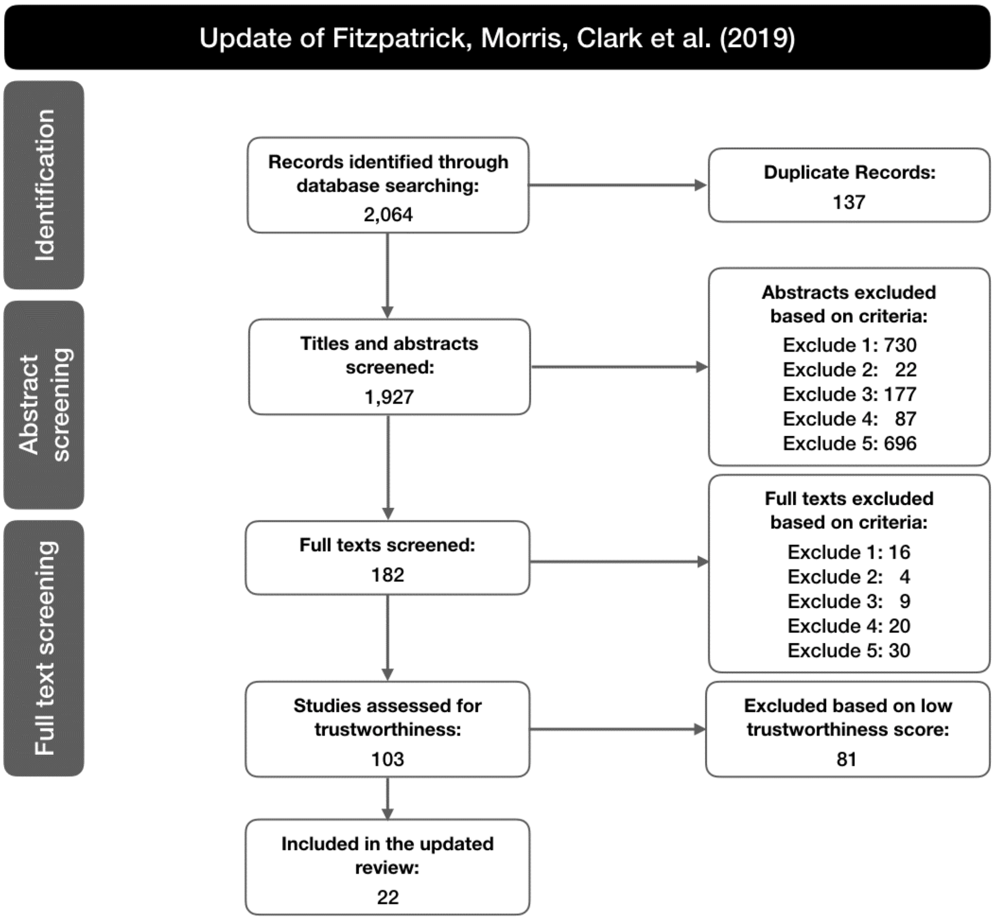
6.5 Appendix 5: CASP scores assigned to systematic reviews during Phase 1

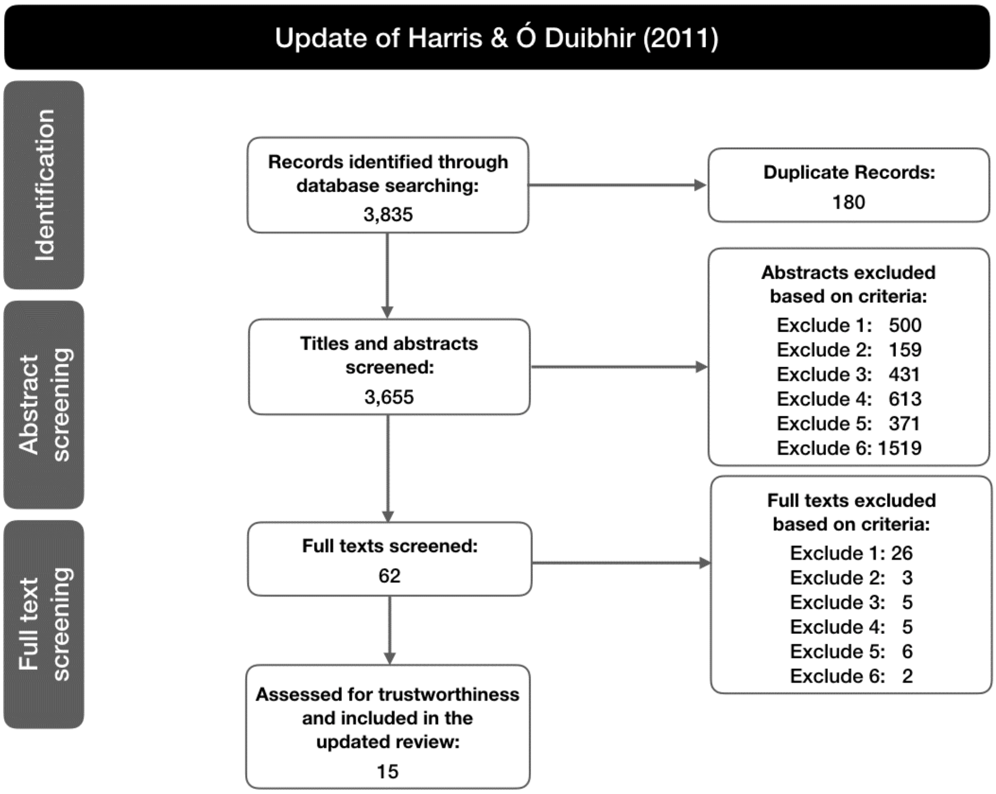
Reference	1.Relevance	2. Focus	3. Sources	4. Methods	5. Inclusiveness	6. Quality	7. Synthesis	8. Clarity	9. Precision	10. Applicability	11. Comprehensiveness	12. Overall Score	Average Score
Alexander (2019)	3	4	4	5	4	3	2	3	0	2	2	2	2.91
Alsadhan (2011)	3	4	5	4	4	0	3	3	4	3	3	3	3.27
Bryfonski & McKay (2019)	5	4	5	4	1	0	3	4	5	5	2	4	3.45
Butler Stewart (2017)	5	4	1	0	0	0	3	3	2	3	4	1	2.27
Chang & Lin (2013)	2	2	2	0	1	0	2	3	3	2	3	2	1.82
Chiu (2013)	3	3	3	1	2	0	1	1	4	2	0	0	1.82
Cole (2013)	3	5	5	2	4	3	3	2	5	3	3	3	3.42
Cole (2014)	2	5	5	3	3	5	5	5	5	2	5	3	4.00
Driscoll et al. (2004)	4	4	3	5	4	3	3	4	0	3	3	3	3.27
Fitzpatrick et al. (2018)	5	5	5	5	5	5	4	5	3	5	5	5	4.75
Fox et al. (2019)	5	4	4	4	3	0	4	3	1	5	3	4	3.27
Goris, Denessen & Verhoeven (2019)	5	4	3	3	4	0	3	4	3	5	3	4	3.36
Graham et al. (2018)	5	4	4	3	3	0	3	3	1	4	3	4	3.00
Grgurović, Chapelle & Shelley (2013)	5	5	5	5	5	5	5	5	5	5	5	4	5.00
Harris & Ó Duibhir (2011)	4	5	4	5	4	3	4	4	2	4	5	4	4.00
Hassan, Macaro, Mason et al. (2005)	4	5	5	5	5	4	4	5	2	4	5	4	4.36
Huang (2016)	5	4	4	4	4	0	4	4	0	5	3	4	3.36
Jeon & Day (2016)	2	5	5	5	5	0	5	5	5	2	5	2	4.00
Jeon & Kaya (2006)	3	4	4	1	2	2	3	3	5	4	3	3	3.09
Kang, Sok & Han (2019)	5	5	5	5	4	0	4	4	5	2	4	4	3.91
Kim & Kim (2017)	5	3	2	3	3	3	3	2	3	2	4	2	3.10
Lee & Huang (2008)	3	4	3	4	3	2	4	4	5	3	3	3	3.45
Lee Jang & Plonsky 2015	2	5	4	5	5	0	5	5	5	2	5	2	3.91
Li (2010)	4	5	5	4	5	3	5	4	5	4	4	4	4.36

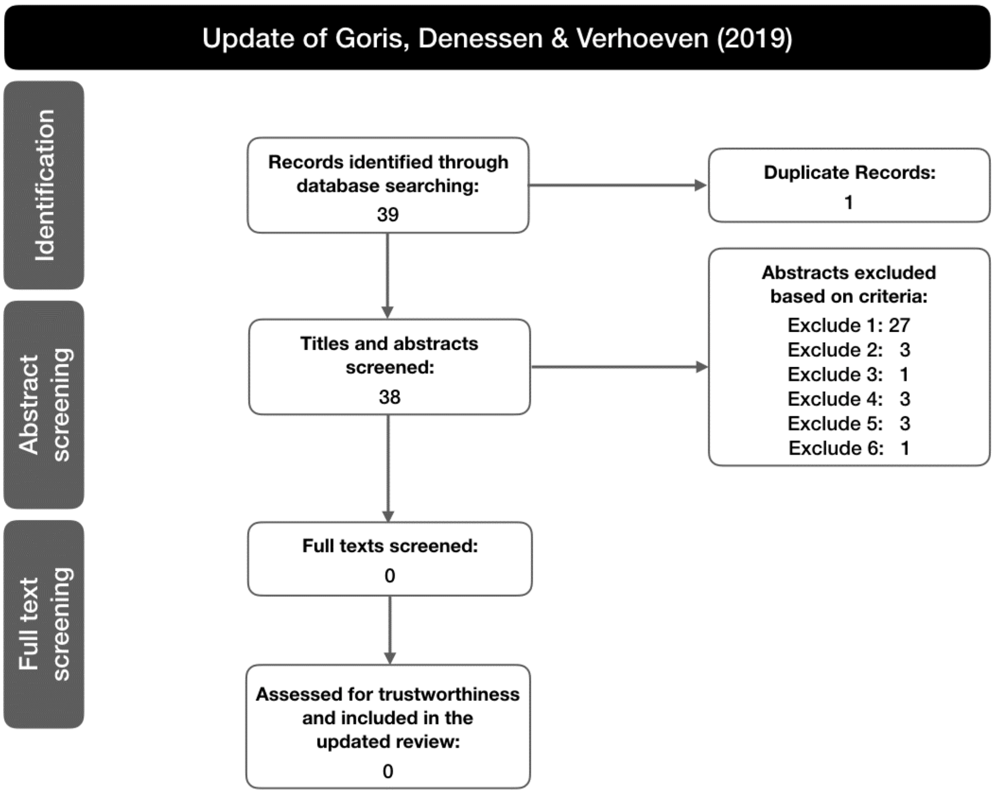
Lin, Huang & Liou (2013)	5	5	5	5	5	5	5	5	5	5	5	5	5.00
Reference	1.Relevance	2. Focus	3. Sources	4. Methods	5. Inclusiveness	6. Quality	7. Synthesis	8. Clarity	9. Precision	10. Applicability	11. Comprehensiveness	12. Overall Score	Average Score
Lo & Lo (2014)	5	5	5	3	4	2	4	5	5	4	4	4	4.18
Lyster & Saito (2010)	3	5	5	4	3	0	5	5	5	3	5	3	3.91
Macaro, Curle, Pun et al. (2018)	5	5	4	4	4	4	5	4	0	4	4	4	3.91
Macaro, Handley & Walter (2012)	5	5	4	5	4	5	5	5	0	5	5	5	4.36
Mahmud (2004)	2	4	2	0	4	0	1	4	2	1	0	0	1.82
Miller (2003)	4	4	5	4	4	3	4	4	3	4	4	4	3.91
Montero Perez et al. (2013)	3	4	5	5	4	1	4	5	5	2	2	2	3.64
Norris & Ortega (2000)	5	5	5	2	2	3	5	4	5	5	4	5	4.09
Norris & Ortega (2001)	5	5	5	1	2	3	4	4	5	5	4	4	3.91
Persson & Nouri (2018)	5	5	3	4	4	0	1	1	0	3	0	2	2.36
Plonsky (2011)	4	5	5	5	5	4	5	5	5	4	5	5	4.73
Rubio-Alcala et al. (2019)	4	3	4	5	5	3	4	3	4	4	4	3	3.91
Russell & Spada (2006)	4	4	3	3	1	1	3	4	5	4	3	3	3.18
Saito (2012)	2	5	5	1	1	0	5	3	5	1	2	1	2.73
Saito & Plonsky (2019)	3	4	5	3	2	2	5	4	5	3	4	3	3.64
Sharifi et al. (2018)	3	5	5	4	5	4	5	4	5	3	3	4	4.18
Shintani (2015)	4	4	5	3	2	2	4	4	5	4	4	4	3.73
Spada & Tomita (2010)	5	4	5	5	4	1	4	4	5	5	2	5	4.00
Sung, Chang & Yang (2015)	3	4	5	5	5	3	5	5	5	4	5	4	4.45
Taj, Sulan, Sipra & Waqar (2016)	5	5	3	0	0	3	4	5	5	5	3	3	3.45
Taylor, Stevens & Asher 2006	5	5	4	4	3	3	3	4	4	4	4	4	3.91
Tsai & Tsai (2018)	3	5	5	5	4	3	5	5	5	4	4	4	4.36
Vahedi, Ghonsooly & Pishghadam (2016)	3	4	3	0	3	2	4	4	4	3	1	2	2.82
Woll & Wei (2019)	4	5	4	3	4	0	3	3	0	2	4	3	2.91
Yun (2011)	3	4	4	4	4	3	4	4	4	3	1	2	3.45

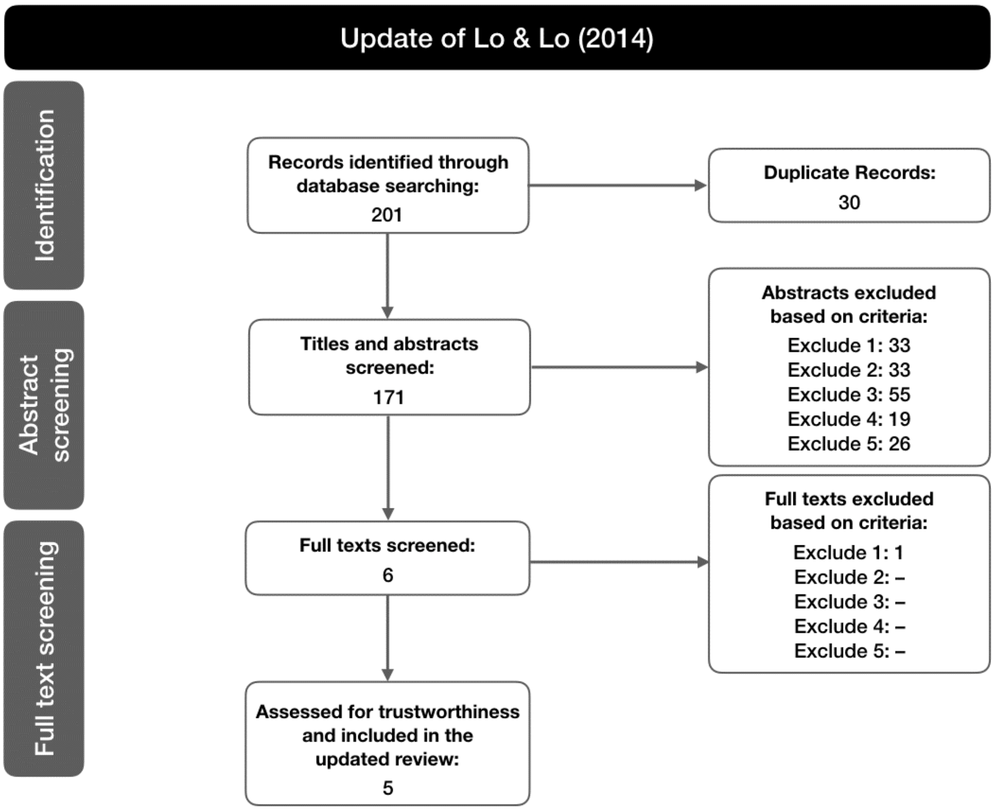
Zhang & Cheung (2018)	2	2	2	2	2	0	2	2	0	1	2	1	1.55

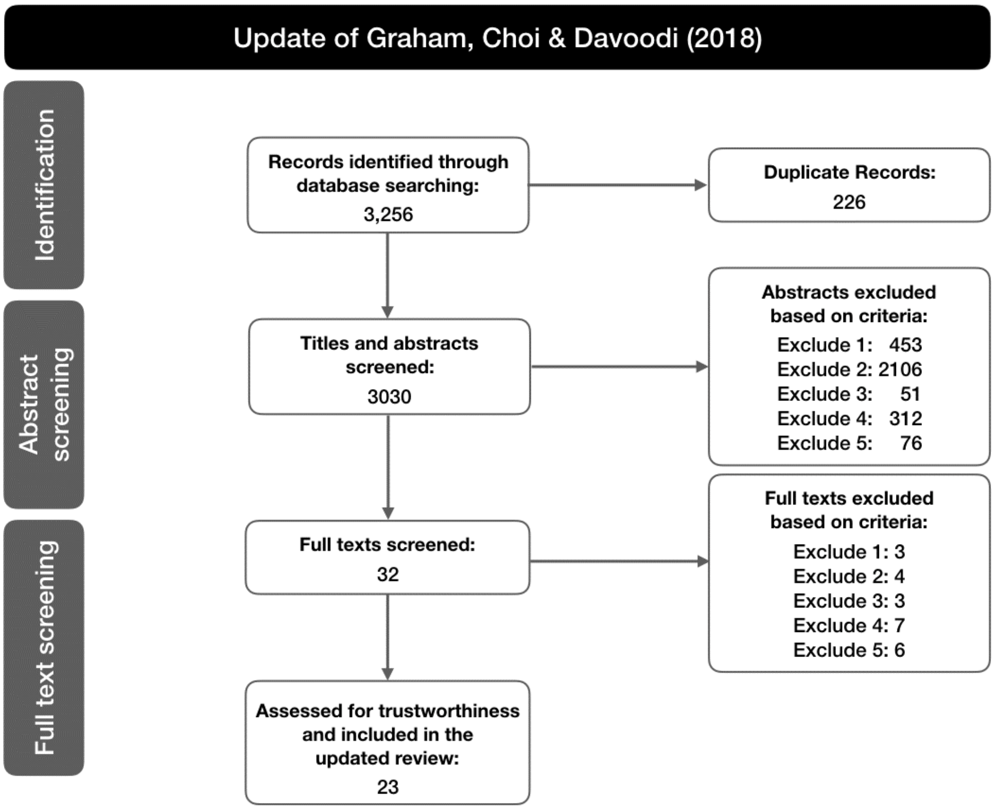
6.6 Appendix 6: Flow diagrams showing the number of studies identified and screened during the updating of the seed reviews


Update of Fitzpatrick, Morris, Clark et al. (2019)
Identification
Abstract screening
Full text screening
Records identified through database searching: 2,064
Duplicate Records: 137
Titles and abstracts screened: 1,927
Abstracts excluded based on criteria:
Exclude 1: 730
Exclude 2: 22
Exclude 3: 177
Exclude 4: 87
Exclude 5: 696
Full texts screened: 182
Full texts excluded based on criteria:
Exclude 1: 16
Exclude 2: 4
Exclude 3: 9
Exclude 4: 20
Exclude 5: 30
Studies assessed for trustworthiness: 103
Excluded based on low trustworthiness score: 81
Included in the updated review: 22


Update of Harris & Ó Duibhir (2011)
Identification
Abstract screening
Full text screening
Records identified through database searching: 3,835
Duplicate Records: 180
Titles and abstracts screened: 3,655
Abstracts excluded based on criteria:
Exclude 1: 500
Exclude 2: 159
Exclude 3: 431
Exclude 4: 613
Exclude 5: 371
Exclude 6: 1519
Full texts screened: 62
Full texts excluded based on criteria:
Exclude 1: 26
Exclude 2: 3
Exclude 3: 5
Exclude 4: 5
Exclude 5: 6
Exclude 6: 2
Assessed for trustworthiness and included in the updated review: 15


Update of Goris, Denessen & Verhoeven (2019)
Identification
Abstract screening
Full text screening
Records identified through database searching: 39
Duplicate Records: 1
Titles and abstracts screened: 38
Abstracts excluded based on criteria:
Exclude 1: 27
Exclude 2: 3
Exclude 3: 1
Exclude 4: 3
Exclude 5: 3
Exclude 6: 1
Full texts screened: 0
Assessed for trustworthiness and included in the updated review: 0


Update of Lo & Lo (2014)
Identification
Abstract screening
Full text screening
Records identified through database searching: 201
Duplicate Records: 30
Titles and abstracts screened: 171
Abstracts excluded based on criteria:
Exclude 1: 33
Exclude 2: 33
Exclude 3: 55
Exclude 4: 19
Exclude 5: 26
Full texts screened: 6
Full texts excluded based on criteria:
Exclude 1: 1
Exclude 2: –
Exclude 3: –
Exclude 4: –
Exclude 5: –
Assessed for trustworthiness and included in the updated review: 5


Update of Graham, Choi & Davoodi (2018)
Identification
Abstract screening
Full text screening
Records identified through database searching: 3,256
Duplicate Records: 226
Titles and abstracts screened: 3030
Abstracts excluded based on criteria:
Exclude 1: 453
Exclude 2: 2106
Exclude 3: 51
Exclude 4: 312
Exclude 5: 76
Full texts screened: 32
Full texts excluded based on criteria:
Exclude 1: 3
Exclude 2: 4
Exclude 3: 3
Exclude 4: 7
Exclude 5: 6
Assessed for trustworthiness and included in the updated review: 23

6.7 Appendix 7: Search strategies per seed review

Databases	Search Terms	Publication dates	Other limitations
FITZPATRICK, MORRIS, CLARK ET AL. (2018)			
- Scopus - LLBA - Web of Science - ERIC ~~- National Library of Wales Catalogue~~ (no access) ~~- Swansea University Library Catalogue~~ (no access)	(child* OR pupil* OR student* OR new speaker*) AND (elementary OR secondary OR high school* OR young OR primary OR early years) AND (second language* OR foreign language* OR modern language* OR heritage language* OR minority language* OR regional language* OR L2) AND (teach* OR learn* OR instruct* OR pedagog* OR acqui*) AND (succe* OR achiev* OR improv* OR attain* OR effect* OR gain* OR increas* OR grow*)	After 2017	- Subject domain: Social Sciences, Arts & Humanities - Publication type: Article, book chapter, article in press, review, book - Language: English, ~~Welsh~~ (no Welsh speakers on the team for this REA) - Peer reviewed only
HARRIS & Ó DUIBHIR (2011)			
~~- Blackwell Synergy/Wiley InterScience~~ (does not exist any more) - Cambridge Journals Online - ERIC - JSTOR - LLBA - Oxford Journals Online - PsychInfo/EBSCO host - Project Muse - Sage Journals	1. (child OR pupil OR student) AND (elementary OR primary OR young) AND (second OR foreign OR modern OR minority OR regional) AND language AND (teach* OR learn* OR pedagogy OR acquisition) AND (effective OR best practice OR best evidence OR success*) 2. (child OR pupil OR student) AND (elementary OR primary OR young) AND (second OR foreign OR modern OR minority OR regional) AND language AND (teach* OR learn* OR pedagogy OR acquisition) AND (effective OR best practice OR best evidence OR success*) 3. (child OR pupil OR student) OR (elementary OR primary OR young) AND (second OR foreign OR modern OR minority OR regional) AND language AND (teach* OR learn* OR pedagogy OR acquisition) AND (effective OR best practice OR best evidence OR success*)	After 2010	- The review team replicated the different limitations Harris and Duibhir (2011) applied to each database search, which are described in detail in their report (p. 118–135).
FOX, CORRETJER & WEBB (2019)			
~~- Communication and Mass Media Complete~~ (no access)	Separate searches for each of the first group of terms in combination with each of the second group.	After June 2019	- Peer reviewed only

- Education Research Complete, incl. ERIC - LLBA - PsychInfo/EBSCO host - Science Direct - Web of SCience	1. foreign language learn*; bilingualism; intercultural competence; multilingualism 2. achievement; academic achievement; aging; attentional control; attitude; auditory competence; benefit; impact; cognitive abilities; cognitive development; cognitive flexibility; cognitive reserve; communicative competence; creativity; enhanced creativity; divergent thinking; empathy; executive function; global awareness; intelligence; intercultural competence; intercultural communicative competence; jobs; career; linguistic process; linguistic reasoning; literacy skills; memory; metalinguistic awareness; motivation; national security; problem-solving; phonological awareness; task switching; verbal abilities; special abilities		
GORIS, DENESSEN & VERHOEVEN (2019)			
- ERIC - PsychInfo/EBSCO host	(CLIL OR bilingual education OR content and language integrated learn*) AND (vocabulary OR grammar OR idioms OR text comprehension)	After February 2018	- Language: English
LO & LO (2014)			
- ERIC - Scopus - Hong Kong Education Bibliographic Database ~~- Education Bureau Central Resources Centre Catalogue~~ (no access)	(medium of instruction OR immersion) AND Hong Kong	After 2009	
GRAHAM, CHOI, DAVOODI ET AL. (2018)			
- ERIC - LLBA - Scopus - PsychInfo/EBSCO host - Web of Science	(English medium instruction OR EMI OR (content and language integrated learning) OR CLIL OR content based instruction OR CBI OR content based language teaching OR CBLT) AND teaching NOT French	After 2017	

6.8 Appendix 8: Inclusion and exclusion criteria per seed review

Fitzpatrick, Morris, Clark et al. (2019): Rapid evidence assessment: Effective second language teaching approaches and methods.

Criterion		
1. Does the publication relate to the effectiveness of something that is identifiable as an approach or method? 'An 'approach' is taken as a set of values, principles, and beliefs about factors that drive learning, and 'method' is taken as the systematic engagement of learners with language.' (p. 10)	Yes, include	No, exclude
2. Does the publication engage with pedagogy in a classroom context?	Yes, include	No, exclude
3. Does the publication have relevance to learners between the ages of three and 16?	Yes, include	No, exclude
4. Does the publication have relevance to teaching of non-dominant target languages?	Yes, include	No, exclude
5. Is the study an RCT or QED? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude

Harris & Ó Duibhir (2011): Effective language teaching: A synthesis of research.

Criterion		
1. Does the study involve learners in the primary school years (ages four–twelve) or inform language teaching for these pupils?	Yes, include	No, exclude
2. Does the study focus on effective language teaching and learning in a school setting within the normal school day?	Yes, include	No, exclude
3. Does the study relate to language teaching and/or learning in one of the following contexts? L2 teaching: Core second language (L2) programmes (and L3 in the case of immigrant children), where the language is taught as a subject L2 immersion: L2 immersion settings, where the second language is the language of instruction for all or part of the school day Heritage language: In heritage/minority/regional/endangered language programmes, where the goal is language maintenance in the case of L1 pupils and language revitalisation in the case of L2 pupils Exclude studies concerned with immigrant L2 learners of English.	Yes, include	No, exclude
4. Does the study have a process-product type design with well-defined independent (effective instructional practices or approaches) and dependent (e.g. pupil performance, or attitudes) variables?	Yes, include	No, exclude

5. Is the study an RCT or QED? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude

Fox, Corretjer & Webb (2019): Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2012–2019.

Criterion		
1. Does the study focus on the benefits of foreign language learning or bilingualism, such as: Aging and health Employability Academic achievement Communicative and intercultural competence Enhanced creativity Cognitive abilities and benefits Including executive functioning/cognitive control, metalinguistic awareness, cognitive development, linguistic processing and reasoning, and spatial reasoning Other	Yes, include	No, exclude
2. Is the study an RCT or QED? Or, in the case of studies concerning bilingualism, was there comparison of bilingual and monolingual participants? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude

Goris, Denessen & Verhoeven (2019): Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies.

Criterion		
1. Does study focus on content and language integrated learning (CLIL)?	Yes, include	No, exclude
2. Is the language of instruction a foreign language? Exclude cases where the language used as the medium of instruction is not a foreign language, e.g. English-Welsh programs in Wales.	Yes, include	No, exclude
3. Does the publication contain a clear measure of one or more FL skill, or of content learning outcomes?	Yes, include	No, exclude
4. Does the study have participants in mainstream primary or secondary education in a European country?	Yes, include	No, exclude

5. Is the study an RCT or QED? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude
6. Is the study longitudinal (has more than one measure in time of the same cohort)?		

Graham, Choi, Davoodi, Razmeh & Dixon (2018): Language and content outcomes of CLIL and EMI: A systematic review.

Criterion		
1. Is the publication a research article? Exclude e.g. book chapters, systematic reviews, meta analyses, and commentaries	Yes, include	No, exclude
2. Is the study focused on teaching content through English (content-based instruction, i.e. EMI/CLIL)? Exclude studies addressing English for Academic Purposes (EAP) and English for Special Purposes (ESP)	Yes, include	No, exclude
3. Was the course instructional language English in a setting where English is not the majority language (i.e. EFL setting)?	Yes, include	No, exclude
4. Does the study directly compare students' learning outcomes in content-based instruction (CBI) and non-content-based instruction (non-CBI) settings?	Yes, include	No, exclude
5. Is the study an RCT or QED? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude

Lo & Lo (2014): A meta-analysis of the effectiveness of English-medium education in Hong Kong.

Criterion		
1. Is the study empirical? Exclude conceptual, theoretical, and review articles	Yes, include	No, exclude
2. Does the study compare students' learning in English-medium (EM) education with those in Chinese-medium (CM) education?	Yes, include	No, exclude
3. Does the study examine students' learning in secondary schools?	Yes, include	No, exclude

4. Does the study have independent student outcome measures, such as achievement, self-concept, or motivation? Exclude studies where the outcomes are teachers' or students' reflections, attitudes, or opinions.	Yes, include	No, exclude
5. Is the study an RCT or QED? Include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. Exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs	Yes, include	No, exclude

6.9 Appendix 9: Phase 2 data extraction sheet

Reference:	
Item	**Data**
Design. Describe the design of the study (e.g. RCT, matched comparison, RDD etc.)	
Participants. Describe in as much detail as is given in the report who took part in the study. Include, for example, age, gender, socio-economic status, L1 etc.	
Intervention. Describe the treatment intervention being evaluated in as much detail as necessary/available to understand what the treatment intervention was.	
Comparator*. Describe the intervention against which the treatment is being compared, in as much detail as necessary/available to understand what the comparison intervention was.	
Outcomes. Describe the outcome measures used in the study.	
Results. Look for effect sizes and confidence intervals, or means and standard deviations, or raw scores. Report the results of statistical tests, if used. If the results are narrative, summarise them here. If results were disaggregated by participant characteristics (e.g. moderator analyses were included in the report) report them here, or append them to the bottom of this sheet.	
Factors contributing to success. Record here any reporting in the study that provides information (or hypotheses) about what contributed to the success of the interventions (assuming they were successful).	
Limitations. Did the authors identify limitations to their study that may compromise the trustworthiness of the findings? This might include attrition, fidelity to the intervention, disruption to the schedule etc. If you have noticed anything that may have compromised the trustworthiness of the findings, which the authors have left unremarked on, note it here.	
Bottom line conclusion. Summarise the bottom line finding of the study. This can be a verbatim excerpt from the text or in the reviewer's own words.	

6.10 Appendix 10: Gorard's Sieve for assessing the trustworthiness of intervention studies

Adapted from: Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics, 110*, 47-59.

Design	Scale	Dropout	Outcomes	Fidelity	Validity	Rating
Fair design for comparison	Large number of cases per comparison group	Minimal attrition, no evidence of impact on findings	Standardized pre-specified independent outcome	Clear intervention, uniform delivery	No evidence of diffusion or other threat	4 ★
Balanced comparison	Medium number of cases per comparison group	Some initial imbalance or attrition	Pre-specified outcome, not standardized or not independent	Clear intervention, unintended variation in delivery	Little evidence of diffusion or other threat	3 ★
Matched comparison	Small number of cases per comparison group	Initial imbalance or moderate attrition	Not pre-specified but valid outcome	Unclear intervention, with variation in delivery	Evidence of experimenter effect, diffusion or other threat	2 ★
Comparison with poor or no equivalence	Very small number of cases per comparison group	Substantial imbalance and/or high attrition	Outcome with issues of validity or appropriateness	Poorly specified intervention	Strong indication of experimenter effect, diffusion or threat	1 ★
No report of comparator	A trivial scale of study, or N unclear	Attrition not reported or too high for any comparison	Too many outcomes, weak measures, or poor reliability	No clearly defined intervention	No consideration of threats to validity	0 ★

As there were no established guidelines as to the cutoffs for each rating in regard to Scale and Dropout, the team agreed upon the following guidelines to establish consistency in ratings for the purpose of this REA:

1. Scale: Number of cases per comparison group
   - Very small: $n \leq 10$
   - Small: $10 < n \leq 25$
   - Medium: $25 < n \leq 50$
   - Large: $n \geq 50$

2. Dropout:
   - Minimal attrition: less than 5% dropout / over 95% completion
   - Low attrition: between 5–15% dropout / between 85–95% completion
   - Moderate attrition: between 15–25% dropout / between 75–85% completion
   - High attrition: over 25% dropout / less than 75% completion

6.11 Appendix 11: Trustworthiness ratings of the studies included in Phase 2, updating of the seed reviews

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Fitzpatrick et al. (2018)							
Akram et al. (2019)	2	2	1	3	1	2	1*
Al Masri et al. (2019)	2	4	0	3	1	3	0
Alghamdy (2019)	1	2	4	1	1	2	1*
Alian et al. (2018)	4	3	4	2	2	2	2*
Al-Jarrah et al. (2018)	2	2	4	1	2	2	1*
Al-Murtadha (2019)	1	4	3	3	4	3	1*
Altin & Saracaloglu (2018)	2	0	0	3	1	3	0
Anjum et al. (2019)	1	2	4	3	1	2	1*
Ansari & Ansari (2018)	2	2	0	3	1	2	0
Awada & Ghaith (2018)	4	2	0	3	4	2	0
Ayçiçek & Yanpar Yelken (2018)	4	2	4	2	1	2	1*
Babapour et al. (2019)	3	2	3	3	2	3	2*
Badawi (2019)	2	3	0	3	3	3	0
Banaruee et al. (2018)	2	2	0	2	3	3	0

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Bataineh & Alqatnani (2019)	2	3	0	3	1	3	0
Bavi (2018)	2	2	0	3	2	0	0
Benitez-Correa et al. (2019)	2	2	4	3	2	3	2*
Bilici et al. (2018)	2	4	0	3	4	3	0
Canado & Perez Canado (2018)	1	4	0	2	2	3	0
Chan (2018)	4	2	0	3	4	3	0
Chan (2019)	3	2	0	3	4	3	0
Chang & Lu (2018)	4	1	0	4	1	3	0
Chen et al. (2018)	4	3	4	4	4	3	3*
Coskun & Eker (2018)	3	3	0	1	4	3	0
Daneshfar et al. (2018)	2	3	4	1	2	1	1*
Doski & Çele (2018)	4	2	4	3	4	3	2*
Duman & Yavuz (2018)	2	3	0	2	1	3	0
Gürergene (2019)	1	2	4	1	1	1	1*
Guerrero et al. (2018)	1	4	0	1	1	0	0
Gurkan & Gurkan (2019)	2	3	0	3	3	3	0
Heller et al. (2019)	3	4	3	4	4	3	3*

Homer et al. (2018)	2	4		3	3	3	2*

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Hsieh & Huang (2019)	2	2	0	1	4	3	0
Iqbal & Rafi (2018)	1	4	0	3	2	3	0
Jalalian (2018)	3	3	0	1	2	2	0
Jelodar & Farvardin (2019)	3	2	3	1	2	2	1*
Kafipour et al. (2018)	2	3	0	3	2	2	0
Karaazmak (2018)	2	3	0	3	2	3	0
Kasprowicz & Marsden (2018)	3	3	4	3	3	4	3*
Kasprowicz et al. (2019)	1	3	1	3	4	3	1*
Keezhatta & Omar (2019)	1	3	4	3	1	2	1*
Kosak-Babuder et al. (2019)	4	4	4	4	4	4	4*
Koukourikou et al. (2018)	2	2	0	4	4	3	0
Lan et al. (2018)	4	2	4	3	4	0	0
Lancaster (2018)	1	4	0	2	2	3	0
Liao et al. (2018)	2	4	4	3	4	3	2*
Lin (2019)	1	2	0	3	4	3	0
Liu, K-P et al. (2018)	3	0	4	3	4	3	0
Liu, M-F et al. (2018)	2	3	0	1	2	2	0

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Ludke (2018)	2	0	2	3	4	3	0
Ma'azi & Janfeshan (2018)	2	2	0	3	4	2	0
Mannion & Griffin (2018)	2	2	0	3	3	1	0
Meguro (2019)	1	2	3	3	2	3	1*
Merino & Lasagabaster (2015)	3	0	3	2	1	4	0
Merino & Lasagabaster (2017)	4	4	4	4	4	4	4*
Meurers et al. (2019)	4	4	4	3	4	3	3*
Meyerhöffer & Dreesmann (2019)	1	3	4	2	2	3	2*
Mirshekaran et al. (2018)	3	3	0	3	0	2	0
Mohaidat (2018)	4	3	4	3	2	3	2*
Mohamadi (2018)	2	3	0	3	4	3	0
Mohammadian (2018)	1	2	4	3	2	2	1*
Moon & Oh (2018)	3	2	4	2	4	2	2*
Mousavi et al. (2018)	4	2	2	2	4	2	2*
Mustapha (2018)	2	3	0	3	1	3	0
Nişanci (2018)	1	3	0	4	2	3	0

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Owen et al. (2019)	2	2	4	3	2	3	2*
Padial-Ruz et al. (2019)	1	3	3	1	2	2	1*
Pamittan (2019)	1	3	0	1	3	2	0
Park et al. (2019)	3	3	4	4	4	3	3*
Park & Oh (2018)	3	3	4	3	2	3	2*
Pfenninger & Singleton (2018)	4	4	0	3	0	3	0
Ponce et al. (2018)	2	2	1	4	3	3	1*
Pujadas & Muñoz (2019)	3	3	3	3	4	4	3*
Rachels & Rockinson-Szapkiw (2018)	2	4	2	3	4	3	2*
Rahayu & Margana (2018)	3	3	0	0	1	0	0
Dewi (2018)	1	3	4	3	1	2	1*
Reynolds & Shih (2019)	1	3	4	4	4	3	1*
Roohani & Rahimi (2018)	2	2	4	3	2	2	2*
Rostamian et al. (2018)	4	2	4	4	4	3	2*
Ruiz de Zarobe & Zenotz (2018)	2	3	0	3	4	3	0
Serrano & Huang (2018)	2	3	3	4	4	3	2*
Shark (2019)	3	3	4	3	1	2	1*

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Shi (2018)	1	2	0	1	3	3	0
Shih (2019)	1	3	4	4	4	4	1*
Silva & Otwinowska (2018)	3	2	0	3	4	3	0
Suárez & Gesa (2019)	3	3	4	3	4	4	3*
Teng (2019a)	4	3	4	3	4	3	3*
Teng (2019b)	3	3	4	3	4	4	3*
Teng (2019c)	3	4	4	3	4	3	3*
Thomas (2018)	2	2	0	0	3	2	0
Ulbricht (2018)	3	3	4	3	3	3	3*
Uzum & Pesen (2019)	2	2	4	2	1	2	1*
Van de Guchte et al. (2019)	3	2	4	3	4	3	2*
van de Ven et al. (2019)	4	3	4	3	3	4	3*
Vyn et al. (2019)	3	3	4	3	4	4	3*
Wang (2019)	2	2	4	4	4	3	2*
Wang et al. (2019)	1	3	4	3	2	2	1*
Winasih et al. (2019)	1	3	4	1	2	2	1*
Yavuz & Arslan (2018)	4	3	4	3	2	3	2*
Yeung et al. (2019)	2	4	0	4	4	3	0
Zhang et al. (2019)	1	2	4	3	4	3	1*

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Fox et al. (2019)							
Antón et al. (2019)	2	4	4	4		4	2*
Comishen et al. (2019)	2	2	0	3		3	0
Damian et al. (2019)	3	2	4	4	4	3	2*
Fecher & Johnson (2018)	2	2	0	3		3	0
Festman & Schwieter (2019)	1	4	4	4		4	1*
Gunzenhauser et al. (2019)	2	3	0	3			0
Kalia et al. (2019)	3	4	4	4	3	2	2*
Lorge & Katsos (2018)	1	2	4	3		2	1*
Marini et al. (2019)	2	3	4	4		4	2*
Paap et al. (2019)	2	4		3			2*
Papageorgiou et al. (2019)	2	3	4	4		4	2*
Paplikar et al. (2019)	2	2	4	3		3	2*
Robinson & Sorace (2019)	2	2	4	4	4	3	2*
Serratrice & De Cat (2018)	2	4	4	4	4	3	2*
Singh et al. (2019)	3	3	2	4		4	2*
Tran et al. (2019)	3	2	4	2	4	3	2*

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Valis et al. (2019)	4	2	4	1	3	1	1*
van Veen et al. (2019)	1	4	4	4		4	1*
Graham et al. (2018)							
Fleckenstein et al. (2019)	2	4	4	3	3	4	2*
Fung & Yip (2014)	3	4	4	4	4	3	3*
Karimi et al. (2019)	4	2	4	3	2	3	2*
Kuzminska et al. (2019)	2	3	0	2	1	1	0
Meyerhöffer & Dreesmann (2019a)	2	4	0	2	2	3	0
Meyerhöffer & Dreesmann (2019b)	3	2	4	3	3	4	2*
Pladevall-Ballester (2018)	2	3	4	3	3	4	2*
Salvador-García et al. (2019)	3	2	4	4	3	4	2*
San Isidro & Lasagabaster (2018)	3	2	4	3	2	4	2*
Harris & Ó Duibhir (2011)							
Aljohani (2016)	1	1	4	1	4	0	0
Alvarez-Marinelli et al. (2016)	4	3	3	4	3	4	3*
Balcı & Çakır	1	1	4	1	4	2	1*
Bavi (2018)	2	2	4	1	1	0	0

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Overall Rating
Berens et al. (2013)	1	1	4	4	4	4	1*
Buckingham & Alpaslan (2017)	1	1	4	2	4	2	1*
Coyle & Roca de Larios (2014)	2	2	2	3	4	1	1*
de Zarobe & Zenotz (2015)	1	1	4	2	1	1	1*
Gutiérrez Martínez & Ruiz de Zarobe (2017)	1	1	4	2	1	1	1*
Murphy (2014)	1	1	4	2	3	2	1*
Nemati et al. (2017)	1	2	0	0	0	1	0
Safataj & Amiryousefi (2016)	1	3	4	1	1	0	0
Shi (2018)	1	1	4	0	1	0	0
Shintani (2011)	1	1	4	1	3	3	1*
Ziegler (2014)	3	4	2	1	1	2	1*
Lo & Lo (2014)							
Hennebry & Gao (2018)	4	4	4	3		3	3*
Fung & Yip (2014)	4	3	4	3	4	3	3*
Lo & Murphy (2010)	2	2	2	4		3	2*
Lau & Yuen (2011)	1	4	4	1		2	1*