The Effects of Using the L1 in Comparison to the L2 to Teach Target Language Vocabulary Items in Primary or Secondary School Settings: A Systematic Review

Lucy Alice Armstrong

MSc in Applied Linguistics for Language Teaching, 2023

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Lucy Alice Armstrong
Date:	14/02/2024

Note that some graphs/tables/images may be removed in order to comply with copyright restrictions.



The Effects of Using the L1 in Comparison to the L2 to Teach Target Language Vocabulary Items in Primary or Secondary School Settings: A Systematic Review

Candidate Number: 1060020
Word Count: 19,997

Lucy Alice Armstrong
Kellogg College

Dissertation Submitted to the Department of Education at the University of Oxford in Partial Fulfilment of the Requirements for an MSc in Applied Linguistics for Language Teaching

Trinity Term, 2023

Table of Contents

Abstract i
List of Tables ii
List of Figures iii
List of Abbreviations iv

Chapter One: Introduction
1.1: Aims of the Review 1
1.2: Structure of the Review 1

Chapter Two: Background and Rationale
2.1: The Role of the L1 in L2 Classrooms 2
2.2: The Monolingual Principle 3
2.3: Prominent Theories of Bilingualism 3
2.3.1. The Notion of a Common Underlying Proficiency 4
2.3.2. The Distinction Between CALP and BICS 5
2.3.3: The Linguistic Interdependence Hypothesis 6
2.3.4: The Threshold Hypothesis 7
2.4: Evidence From Correlational Research 9
2.4.1: Correlations Between L1 and L2 Phonological Awareness 10
2.4.2: Correlations Between L1 and L2 Knowledge and Skills 11
2.4.3: Correlations Between Dissimilar Languages 13
2.4.4: The Limitations of Correlational Research 14
2.5: Evidence From Bilingual Education Programmes 14
2.5.1: Evidence From Randomised Studies 16
2.5.2: The Implication of This Research for L2 Teaching 16
2.6: How Multilingual Learners Use Their L1s 17
2.7: Recommendations for Pedagogical L1 Use 18
2.7.1: Pedagogical Translanguaging 18
2.7.2: The Limitations of These Practices 19
2.8: Experimental Research on the Effects of L1 Use 19
2.8.1: Using the L1 to Teach L2 Vocabulary 20

Chapter Three: Methodology
3.1: The Research Question 21
3.2: Eligibility Criteria 22
3.3: Information Sources 24
3.4: The Search Strategy 25
3.5: The Selection Process 26
3.5.1: Title/Abstract Screening 26
3.5.2: Full-text Screening 26
3.6: The Data Extraction Process 27
3.7: Risk of Bias Assessment 27
3.8: Synthesis of Results 29

Chapter Four: Results
4.1: Study Selection 29
4.2: Study Characteristics 31
4.2.1: Publication Date and Type 31
4.2.2: Language Programme Type and Location 32
4.2.3: Study Design and Focus 33
4.2.4: Study Size and Duration 34
4.2.5: Participant Proficiency, Gender and L1 Background 35
4.2.6: Outcome Measures and Study Findings 36
4.3: Risk of Bias Results 36
4.3.1: Selection Bias 39
4.3.2: Study Design 39
4.3.3: Confounders 40
4.3.4: Blinding 40
4.3.5: Data Collection Methods 40
4.3.6: Withdrawals and Drop-Outs 41
4.4: Limitations of the Quality Assessment 41

Chapter Five: Discussion
5.1: Using L1 Translations or Pictures 42
5.2: Watching Television with L1 or L2 Audio and Subtitles 44
5.2.1: Studies on Incidental Vocabulary Learning 45
5.2.2: Studies on Intentional Vocabulary Learning 46
5.3: Listening with L1 or L2 Glossing, Storytelling or Contrastive Analysis 47
5.3.1: Studies on Oral or Written Glossing 48
5.3.2: Studies on Oral Storytelling 49
5.3.3: Studies on Contrastive Analysis 50
5.4: Reading with L1 or L2 Glossing or Writing Activities 50
5.4.1: Studies on Written or Oral Glossing 51
5.4.2: Studies on Written and Oral Glossing 52
5.4.3: Studies on Written Glossing and Writing Activities 53
5.5: Reading with Translation Activities, Content Discussion or Vocabulary Tasks 54
5.6: Using Translation-Based or Textbook-Based Learning Activities 54

Chapter Six: Conclusion
6.1: Summary of the Findings 55
6.2: Limitations of the Review 56
6.3: Recommendations for Future Research 57

Reference List 58

Appendices 77

Abstract

There has long been debate about what role – if any – the first language (L1) ought to assume in second or foreign language (L2) classrooms, and following the multilingual turn in language teaching, recent pedagogical advice has seen its use widely advocated. Generally, theoretical justifications are offered in support of this while empirical justifications are lacking, however, and so this review aims to remedy this by establishing the nature and extent of experimental research relevant to these claims. It builds on an earlier systematic review by Chalmers (2019) into the educational effects of using L1-mediated teaching approaches in L2 classrooms with primary and pre-primary learners by narrowing the focus to L1-mediated vocabulary teaching approaches in particular, and it also expands the range of L2 learners examined by identifying research at both the primary and secondary levels. Most importantly, it requires L1-mediated vocabulary teaching approaches to be directly compared to L2-mediated vocabulary teaching approaches, and for outcomes to be quantitative measures of L2 vocabulary acquisition.

In addition, all of the included studies must record intervention research with typical student populations in formal educational environments, and have been published between 1980 and 2023. Altogether, systematic searches of five online databases uncovered 45 studies that met these inclusion criteria, and although the vocabulary teaching interventions that they explored varied, they can be broadly defined as having investigated six key types. Within these, studies tended to favour L1 use – thereby providing support for recommendations to include it in L2 lessons – but this support is tentative on account of the studies (1) having a generally high risk of bias, (2) being relatively small-scale, and (3) being interpreted by only one reviewer. There were also several that produced conflicting findings, while seven that were selected for Stage two screening could not be located, and so the need for additional and more readily available research is clear. Finally, this review also highlighted the ubiquity of ambiguous or incomplete reporting in education research, and the lack of studies using individual randomisation.

List of Tables

Table 1: Summary of the Five Studies on L2 Vocabulary Teaching 20

Table 2: The Inclusion and Exclusion Criteria 22

Table 3: The Data Extracted from Included Studies 27

Table 4: Summary of the EPHPP Quality Assessment Criteria 28

Table 5: The Included Studies by Publication Type 32

Table 6: The Included Studies by Country 32

Table 7: The Structure of the Included Studies 33

Table 8: The Focus of the Included Studies 34

Table 9: The Size of the Included Studies 35

Table 10: The Duration of the Included Studies 35

Table 11: The Results of the Risk of Bias Assessment 37

Table 12: Summary of the Studies in Section 5.1 42

Table 13: Summary of the Studies in Section 5.2 44

Table 14: Summary of the Studies in Section 5.3 47

Table 15: Summary of the Studies in Section 5.4 50

Table 16: Summary of the Studies in Section 5.5 54

Table 17: Summary of the Studies in Section 5.6 54

Table 18: Summary of Findings by Intervention Type 55

List of Figures

Figure 1: The Separate Underlying Proficiency Model of Bilingualism 4

Figure 2: The Common Underlying Proficiency Model of Bilingualism 5

Figure 3: The 'Dual Iceberg' Representation of Bilingual Proficiency 6

Figure 4: The Types, Levels and Effects of Bilingual Competence 8

Figure 5: The PRISMA Flow Diagram 30

Figure 6: The Included Studies by Publication Year 31

List of Abbreviations

The following abbreviations and acronyms are used throughout this dissertation:

BICS	Basic Interpersonal Communicative Skills
CALP	Cognitive/Academic Language Proficiency
CEFR	Common European Framework of Reference
CUP	Common Underlying Proficiency
EAL	English as an Additional Language
EFL	English as a Foreign Language
ELT	English Language Teaching
ESL	English as a Second Language
FFL	French as a Foreign Language
FSL	French as a Second Language
GFL	German as a Foreign Language
IFL	Italian as a Foreign Language
L1	First language
L2	Second language
NRC	Non-Randomised Comparison
RCT	Randomised Controlled Trial
SFL	Spanish as a Foreign Language
SUP	Separate Underlying Proficiency

Chapter One: Introduction

1.1: Aims of the Review

The question of how vocabulary can most effectively be acquired by second/foreign language (L2) learners has been integral to successful L2 learning since the practice first began (Nation, 2001). Overall, research into the issue has tended to focus on the efficacy of intentional vocabulary learning in comparison to incidental vocabulary learning (Read, 2000), and the variety of practices that can be employed in relation to these two approaches (e.g. Webb et al., 2020, 2023). According to Hulstijn (2001), intentional vocabulary learning entails the use of techniques, resources or tasks that are designed to commit novel lexical items to long-term memory. Incidental vocabulary learning, on the other hand, occurs when such items are inadvertently acquired as the by-product of activities with other objectives, and while both forms of acquisition are effective (Webb & Nation, 2017), a fundamental concern is whether they are best facilitated through the students' first language (L1), or through their L2.

Empirical research that directly compares the two approaches naturally provides the greatest insight into this issue, and so this review will systematically locate, assess and synthesise all such research. It will focus on intervention studies, of any design, that have been conducted with typical student populations in formal educational environments, and, for reasons that will later be outlined, it will be limited to primary and secondary school-aged subjects. Given the purpose of the review, the study outcomes will be quantitative measures of L2 vocabulary acquisition – be it intentional or incidental – and the collated evidence will be used to identify the effects of L1-mediated vocabulary teaching in comparison to L2-mediated vocabulary teaching. Whether one consistently outperforms the other, or whether they fare variously under different learning conditions, will be critically evaluated, and conclusions as to the appropriacy of the two approaches from a pedagogical perspective will be presented.

1.2: Structure of the Review

In total, this review consists of six chapters, with Chapter One introducing its overall aim and structure before Chapter Two then provides a more comprehensive account of its rationale and research background. It does this by reviewing the history of L1 use in L2 classrooms, and outlining both theoretical and empirical justifications for the suggestion that L1-mediated

vocabulary teaching may be advantageous. Following this, Chapter Three identifies the research question of the review, and provides a thorough description of its methodology, from the identification of eligibility criteria, to the data collection process of systematically locating, screening and selecting empirical studies for inclusion. How these studies were assessed for their quality is also described, and Chapter Four presents the results of the review in terms of the studies identified, and their defining characteristics. In Chapter Five, a more detailed analysis of the nature and findings of the research is presented, with inferences draws as to the efficacy of different teaching approaches, and finally, Chapter Six provides a summary of the key conclusions, limitations, and recommendations for future research of the review.

Chapter Two: Background and Rationale

2.1: The Role of the L1 in L2 Classrooms

Historically, the degree to which the L1 of L2 learners has served as a resource in L2 classrooms has varied according to the pedagogical approach employed by teachers and academic institutions at any given time (Cook, 2001). Preceding the Great Reform in language teaching towards the end of the nineteenth Century, for example – wherein L2 classrooms sought to replicate the conditions of monolingual L1 acquisition by using only the target language – the highly popular Grammar Translation method relied on its use (Hawkins, 1987). Primarily, this was in order to teach the dead languages of Latin and Ancient Greek through translation, with the L1 dominating classroom instruction (Larsen-Freeman, 2011), but come the development of the successive Direct Method, the prevalence L1 use diminished (Bayley, 1998). This is because it was founded on the monolingual principle of exclusive L2 use, and as this method proved better suited to the teaching of modern languages (Howatt & Widdowson, 2004), its assumptions have continued to inform L2 teaching methodologies since the 1880s.

Importantly, this does not mean that the L1 ceased to be utilised at all, with the contrastive analysis hypothesis by Lado (1957) proposing that teachers compare it to the target language in order to identify areas of L2 learning ease or difficulty. Following this, they could then adapt their pedagogical approach accordingly, but of course, advocating for a consideration of the L1 so as to prevent – or at least alleviate – interference with the L2 is not the same as actively endorsing its use within the classroom. In fact, the suggestion that its areas of dissimilarly may

be responsible for negative linguistic transfer – which derived from a behaviourist view of language use as habit formation (Mitchell & Myles, 2004) – served to reinforce the theory that it should not be used. This view persisted even after L2 errors came to be understood in developmental terms in the 1970s (Benson, 2002), and the input hypothesis by Krashen (1985) further supported it by asserting that L2 learning depends on interlanguage + 1 level input. In short, while the L1 of L2 learners was acknowledged by teachers, the pedagogical approaches of the 20th and early 21st Centuries were dominated by L2-only classroom policies.

2.2: The Monolingual Principle

When reflecting on the logic of the monolingual principle, Cook (2001) describes theorists as supposing that just as L1 acquisition occurs without appeal to another language, so too should L2 acquisition. Not only this, but in addition to the association of the L1 with the outdated Grammar Translation method, and the impact of Krashen's (1985) input hypothesis, Atkinson (1987) found further justification in the facts that monolingual teachers cannot understand the L1 of their students, and L2s cannot be acquired without practicing them. Littlewood and Yu (2011) also speculate that a L2-only stance permeates Western teaching methodologies because it maintains the status of the native-speaker teacher – thereby asserting neo-colonial control (Auerbach, 1993) – and for Atkinson (1987), the consequence of this is that the L1 is a "neglected resource" (241). Others (Butzkamm, 2003, 2007; Harbord, 1992; Nation, 2003) have echoed this claim, and Cummins (2007) states that although language policy and practice appear to regard the monolingual principle as self-evident, it in fact does not reflect prominent theories of the bilingual mind, or how it is that additional languages are learned. In the next section of the chapter, therefore, the most influential of these theories will be reviewed.

2.3: Prominent Theories of Bilingualism

According to Baker (2006), historic perceptions of bilingualism tended to adopt a deficiency perspective, and to regard the existence of two or more languages in the mind as placing an undue burden on the brain. In part, this was a result of the erroneous belief that the capacity of the brain is finite, and thus that attempting to acquire a L2 detracts from the cognitive resources available to the L1 (Chalmers & Murphy, 2022). Cognitive deficiencies such as low intelligence and mental confusion were presumed to follow, and the balance effect hypothesis

by Macnamara (1966) encapsulated this view by proposing that as L2 proficiency increases, L1 proficiency decreases. Sometime later, Cummins (1980a) represented it visually by depicting two balloons inside a human head (see Figure 1), and he referred to this conceptualisation as the Separate Underlying Proficiency (SUP) model of bilingualism. Overall, his intent was to indicate that not only do the two balloons – signifying two languages – compete for space, but they are also entirely separate, such that stimulation of one does not affect the other.

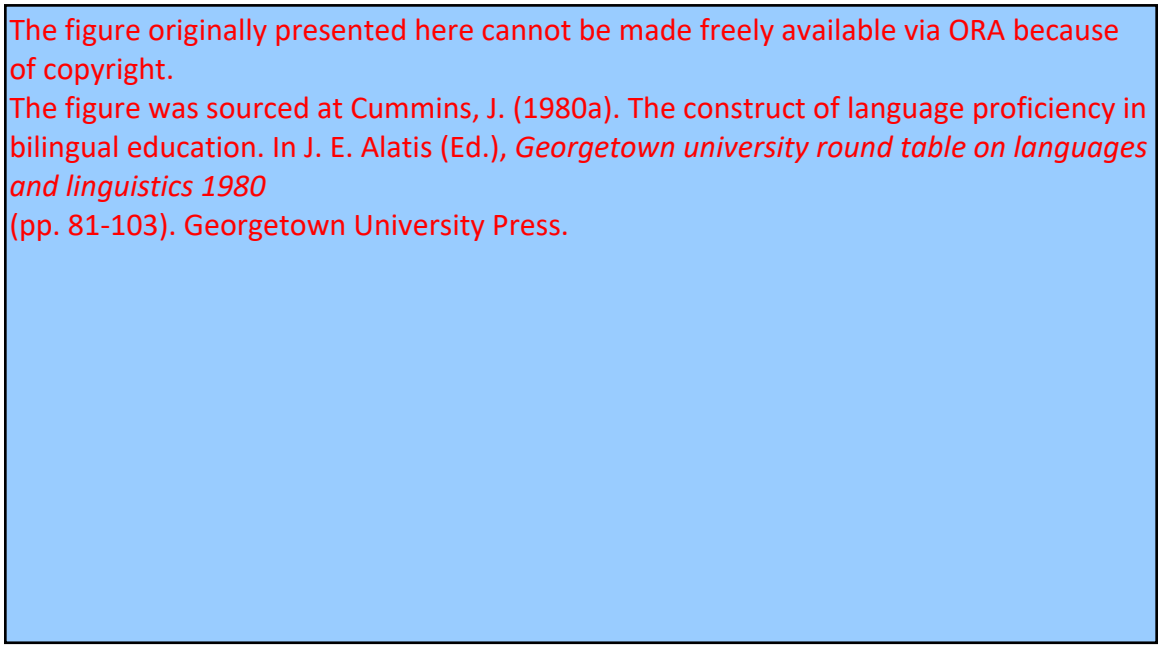


Figure 1: The Separate Underlying Proficiency Model of Bilingualism

2.3.1: The Notion of a Common Underlying Proficiency

For Cummins (1980a), the SUP model of bilingualism signifies a complete misrepresentation of the bilingual mind, despite being constructed on what he concedes are "common sense assumptions" (92). As an alternative, therefore, he put forth the Common Underlying Proficiency (CUP) model of bilingualism (see Figure 2), which represents L1 skills and L2 skills by means of a single balloon with two separate necks – one termed the 'L1 channel', and the other the 'L2 channel'. His theory was that instruction received through either of these channels contributes to the creation of a CUP that is itself comprised of Cognitive/Academic Language Proficiency (CALP) in both languages, and is unlimited in capacity. Furthermore, while this is true for bilingual individuals of any age, he claimed this proficiency to influence the development of young bilingual's literacy skills – and their academic achievement more broadly – to a greater degree than their L1 knowledge or L2 knowledge individually.

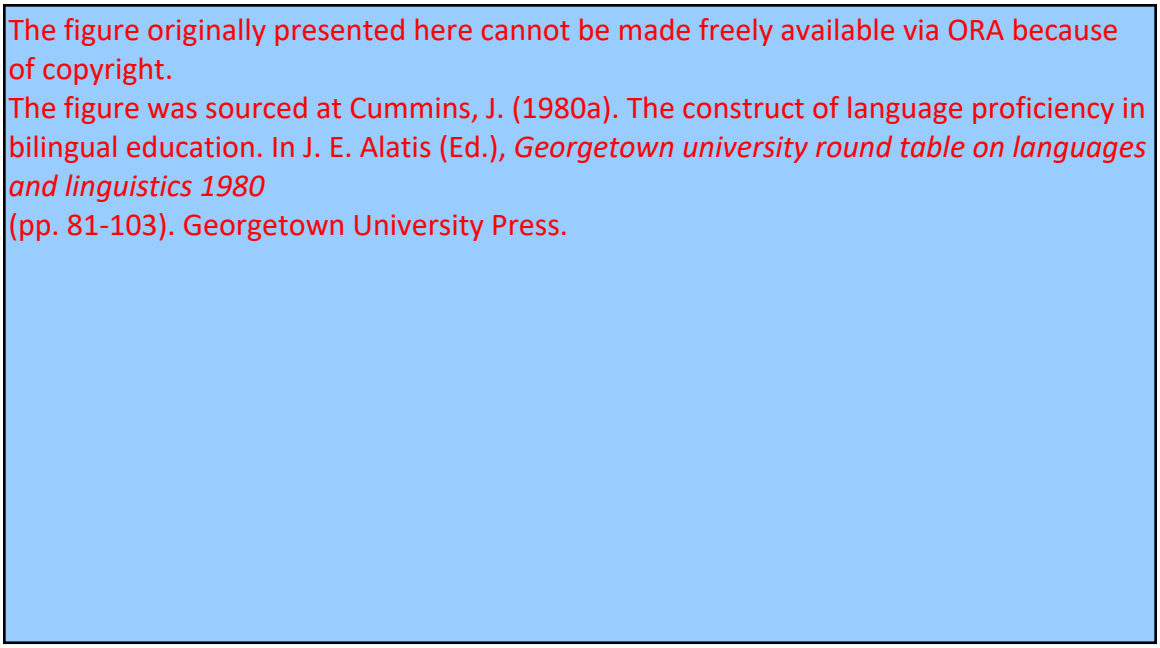


Figure 2: The Common Underlying Proficiency Model of Bilingualism

2.3.2: The Distinction Between CALP and BICS

To elaborate on why Cummins (1980a) proposed this, it is necessary to consider the distinction that he drew between CALP on the one hand, and Basic Interpersonal Communicative Skills (BICS) on the other. Basically, while the term CALP refers to high-level linguistic knowledge that enables learners to make sense of complex educational content, and to engage in decontextualised and often literacy-related cognitive or academic tasks, the term BICS refers to everyday social or transactional language. This language is primarily oral, with the vast majority of individuals acquiring it regardless of their IQ or academic aptitude, and Cummins (1981a; 2000) attributes this to the fact that it is usually context-embedded and cognitively undemanding, which is entirely the opposite of CALP. Unsurprisingly, the latter tends to require instruction, and Cummins (1981a) explains that while the surface-level proficiencies in vocabulary, grammar and pronunciation that constitute BICS are visible in daily interaction, the CALP required to succeed in more challenging academic situations is harder to see.

Regarding support for the BICS-CALP distinction, a corpus analysis by Biber (1986) classified more than a million words of written and spoken English into similar categories, and analogous classifications were also put forward by other theorists. Halliday (1975), for example, differentiated between the pragmatic and mathetic functions of language, Oller (1977) between the social and ideational functions, and Burt and Dulay (1978) between the language

produced by 'natural communication' tasks and 'linguistic manipulation' tasks. Empirical evidence was also found in a study by Skutnabb-Kangas and Toukomaa (1976) when language tests revealed that the surface-level fluency of Finnish migrant children in Swedish was not reflected in their cognitive/academic proficiency. What this indicates is that BICS may be acquired before CALP in immersion contexts, but Cummins (1999) clarifies that the opposite can also happen, and that the two – although conceptually distinct – are interrelated.

2.3.3: The Linguistic Interdependence Hypothesis

Turning to the significance of these theoretical claims for the role of the L1 in L2 teaching, Cummins (1980a; 1981a) asserted that aspects of CALP acquired through one language do not need to be reacquired through another. In other words, they are transferable, with Cummins (2000) giving the example of an individual who understands how to tell the time in their L1 not needing to relearn the concept of time telling in their L2. Rather, they need only to attain the L2 lexis and structures necessary for doing this, and to illustrate this mutually supportive relationship, Cummins (1980a) created the dual iceberg representation of bilingual proficiency (see Figure 3). The purpose of this representation was to emphasise that, despite the surface features of any L1 and L2 rarely being transferable – and perhaps some of their non-surface features also being non-transferable – CALP is interdependent, and contributes to a bilingual individual's performance on cognitive/academic tasks undertaken in either language.

The figure originally presented here cannot be made freely available via ORA because of copyright.
The figure was sourced at Cummins, J. (1980a). The construct of language proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown university round table on languages and linguistics 1980*
(pp. 81-103). Georgetown University Press.

Figure 3: The 'Dual Iceberg' Representation of Bilingual Proficiency

Cummins (1979a; 1981a) called this claim the linguistic interdependence hypothesis, and he made clear that the interdependence of L1 CALP and L2 CALP does not exist in an affective or experiential vacuum. To express this, the official formulation of his hypothesis was:

> To the extent that instruction in Lx is effective in promoting cognitive/academic proficiency in Lx, transfer of this proficiency to Ly will occur provided there is adequate exposure to Ly (either in school or the environment), and adequate motivation to learn Ly. (Cummins, 1980a: 90)

In essence, this means that suitably exposed and motivated L2 learners with well-developed L1 CALP will acquire L2 CALP more quickly, and notably, this coheres with the fact that studies relating age to L2 learning have consistently shown an advantage for older learners (Burstall et al., 1974; Ekstrand, 1977; Ervin-Tripp, 1974; Fathman, 1975; Snow & Hoefnagel-Hohle, 1978). Generally, this advantage manifests in their acquisition of L2 syntax, morphology and literacy skills, and although a study by Ramsey and Wright (1974) appeared to contradict this, a re-examination of its data by Cummins (1981b) shed new light on its interpretation.

Basically, the study reported that migrant children who arrived in Canada below the age of seven demonstrated grade-appropriate English language skills, while those who arrived after that age did not. In fact, they exhibited a negative relationship between their age on arrival, and standardised measures of their English proficiency, but problematically, Ramsey and Wright (1974) did not consider the children's length of residence (LOR) in Canada, or the fact that grade norms become progressively more difficult to meet as students age. It made sense, therefore, to compare their raw scores on the proficiency tests, and upon doing this, Cummins (1981b) discovered that older learners acquired more L2 knowledge in equivalent LORs than younger learners did. In addition, he also found that LOR exerted a significant effect on English proficiency for at least five years, the implication being that L2 CALP takes around five years to acquire in immersion contexts, while L2 BICS takes only two years (Cummins, 2016).

2.3.4: The Threshold Hypothesis

Of course, if high levels of L1 CALP serve to hasten the acquisition of L2 CALP, and low levels to slow it down, then non-existent L1 CALP – coupled with a lack of high quality or sustained L2 instruction – may prevent its development altogether. In such cases, bilingual individuals would possess L1 and L2 BICS, but nothing more, and this possibility led Cummins (1979a) to build on the work of Toukomaa and Skutnabb-Kangas (1976; 1977) by devising the threshold hypothesis. Essentially, this asserts that L2 learners may develop additive bilingualism, where

they have advanced proficiency in two languages, dominant bilingualism, where they have advanced proficiency in only one, or semilingualism, where they have limited proficiency in both (see Figure 4). Which they acquire naturally has consequences for their cognitive and academic development, and Cummins (1979a) suggests that there are two threshold levels of bilingual competence – the lower threshold, whose attainment avoids any potentially negative effects, and the higher threshold, whose attainment facilitates rapid cognitive growth.

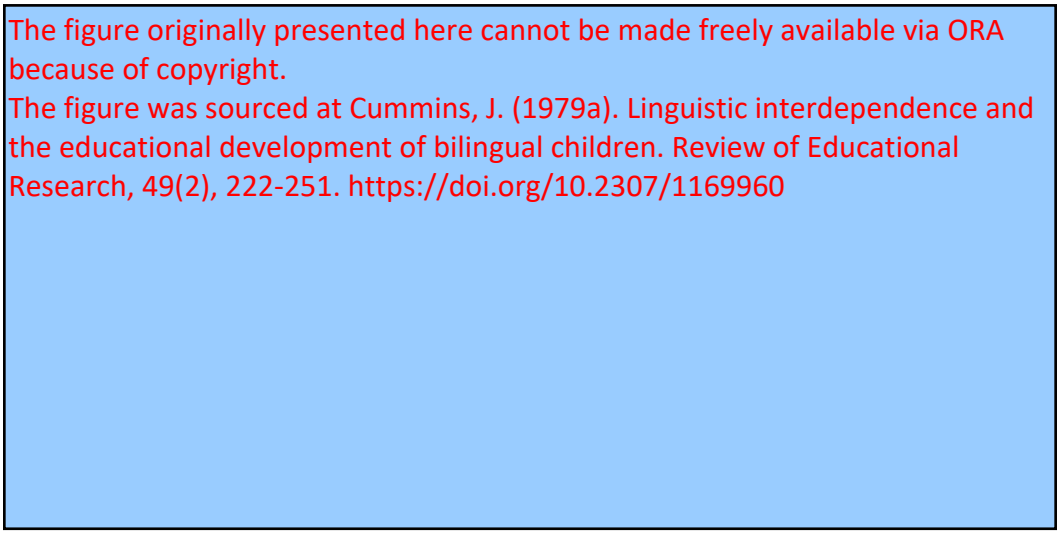


Figure 4: The Types, Levels and Effects of Bilingual Competence

Clearly, a prerequisite for surpassing the higher threshold – and therefore achieving additive bilingualism – is advanced L1 skills, and it is worth noting that these claims are not merely theoretical. Rather, they constitute an attempt to reconcile the conflicting empirical data of multiple studies (see Cummins, 1976 for a review) wherein bilingualism was variously shown to yield positive and negative cognitive effects. Admittedly, the term 'semilingualism' has been widely condemned for decades now (Edelsky et al. 1983; MacSwan, 2000; Martin-Jones & Romaine, 1986) – primarily due to its early associations with the Nazi-propagated notion of pathological bilingualism (Karlander & Salö, 2023), and the language practices of linguistic minority communities (Hansegård, 1968) – and Cummins himself (1979b) conceded that it is problematic. Ultimately, its pejorative connotations and imprecise operationalisation even led him to abandon the term, but not before emphasising that although it "confuses rather than clarifies" (104-105), the phenomenon of low L1 and L2 CALP persists (Cummins, 2000).

2.4: Evidence From Correlational Research

The implication of the linguistic interdependence hypothesis and the threshold hypothesis is that an individual's L1 ought to be developed alongside their L2, so that the two can inform and support one another. Furthermore, if the linguistic interdependence hypothesis is valid, then L1 proficiency can be expected to correlate with L2 proficiency, as well as with other variables of verbal and non-verbal capability (Cummins, 1979b). This correlation will not be absolute, due to linguistic development being time-consuming, non-linear and mediated by motivation and language exposure, but some relationship is predicted, and this leads to the questions of what linguistic proficiency is, and how it can be measured. Correlations can only be calculated if the construct is sufficiently quantifiable, after all, and there has been debate on this issue, with Hemàndez-Chávez et al. (1978) proposing a three dimensional matrix of sixty-four language proficiencies – each of which can be independently assessed – while Oller (1978) stated that a 'global language factor' instead comprises the core of proficiency.

When justifying the existence of this factor, Oller (1978) claimed that it explains the majority of the variance in a wide variety of proficiency measures administered to a range of subjects, with different ages, L1s and educational contexts. It could be evaluated equally well by tasks in any of the four skills, and had defied attempts to isolate its subcomponents (see Anderson, 1976; Oller, 1976; Oller & Perkins, 1978; Scholz et al., 1980). For Cummins (1980b), it was a more practical conceptualisation of proficiency than that provided by Hemàndez-Chávez et al. (1978) – if not necessarily a more accurate one – but he took issue with its one dimensionality, given his distinction between BICS and CALP. For him, CALP was the more significant aspect of proficiency to consider both across languages, and in relation to academic success, and so he advocated for it to be investigated by correlational research. Following this, he then compiled several studies that did this (see Cummins, 1979b), and concluded that they found moderate to strong positive correlations between the L1 CALP and L2 CALP of bilingual children.

Overall, these correlations ranged from 0.42 to 0.77, with CALP generally being assessed by standardised measures of listening or reading comprehension, and cloze tests – both written and oral – that entailed linguistic manipulation. In addition, these studies (Carey & Cummins, 1979; Cummins, 1977; Genesee, 1979; Genesee & Hamayan, 1980; Lapkin & Swain, 1977; Swain et al., 1976; Taft & Bodi, 1980) also revealed that L1 CALP and L2 CALP correlated with

measures of verbal IQ, non-verbal IQ and language learning aptitude, and Cummins (1979b; 1980b) regarded this as supporting his CUP model of bilingualism. What this means, as stated previously, is that cognitive/academic proficiency acquired through a L1 automatically aids L2 learning, but these studies are somewhat dated, and additional empirical research has since been conducted. Whether this research similarly upholds these claims will thus be discussed in the following sections, with a review first of studies that examine the relationship between L1 and L2 phonological awareness, and then that between L1 and L2 knowledge or skills.

2.4.1: Correlations Between L1 and L2 Phonological Awareness

To begin by clarifying the term 'phonological awareness', McBride-Chang (1995) defines it as an individual's knowledge of the sound structure of a language, and it has consistently been associated with L1 word recognition and reading abilities (see Gottardo et al., 1996; Hulme et al., 2002; Share, 1995). In an early attempt to ascertain whether such awareness might also benefit L2 skills, Durgunoglu et al. (1993) administered multiple tests to 31 Spanish-speaking first-grade English students on a bilingual programme, and found that their L1 phonological awareness scores were positively correlated with not only their L1 word and pseudoword recognition scores, but also their L2 ones, at r=0.68 and r=0.51 respectively. In fact, this latter effect was most distinct for pseudowords whose pronunciation would be different in Spanish, which implies that L1 phonological awareness facilitates L2 phonological awareness, and these findings were later supported in more comprehensive research by Comeau et al. (1999).

In this study, 112 English-speaking students of French, from grades 1, 3 and 5, were tested on both their L1 and L2 phonological awareness, with analysis demonstrating that the former was as related to word decoding skills in each language as the latter was. In other words, L1 and L2 phonological awareness were correlated, and this strongly suggests that such awareness is a fundamental cognitive mechanism – and therefore an aspect of CALP – that transfers across languages. Of course, one issue with this inference is that the equivalency of L1 and L2 skills in this experiment may have been due to the students receiving similar training in French and English – especially as they attended a bilingual school. Subsequent studies, however (Quiroga et al., 2002; Dickinson et al., 2004; Lopez & Greenfield, 2004), have repeated this finding in a range of contexts, and Atwill et al. (2007) even furthered it by dividing their 68 kindergarten children into groups of high and low L1 proficiency, and assessing them independently.

How they gauged this proficiency was by means of a receptive vocabulary test, and what they found when they compared the two groups was that no evidence of phonological awareness transferring across languages emerged for the children in the low proficiency group. For those in the high proficiency group, however, positive correlations occurred between their L1 and L2 phonemic awareness scores on two different tests – at r=0.49 and r=0.75 – so in addition to the previous studies reinforcing the linguistic interdependence hypothesis, this study also provides tentative support for the threshold hypothesis. Longitudinal research by Sparks et al. (2008) then determined that evaluations of L1 phonological awareness in 54 English-speaking students in elementary school were predictive of their L2 French, German or Spanish spelling ability in high school, which speaks to the long-term transfer of this awareness if interventions are unable to improve it, and crucially, this transfer is not always from the L1 to the L2.

Research by Murphy et al. (2014) provided perhaps the best example of this when it showed that English-speaking Year 3 children learning L2 French or Italian for 15 weeks outperformed those learning neither language on several post-test measures of L1 literacy skills. These were single-word reading accuracy and three aspects of phonological processing for the students of Italian – namely, alliteration detection, rhyme detection, and spoonerism creation, where the initial sounds or letters of two or more words are successfully transposed. For the students of French, however, this was alliteration detection only, which led the researchers to speculate that learning a language with a shallow orthography, like Italian, might be more beneficial for individuals whose L1 has a deep orthography, like English. This may be true only for languages with alphabetic writing systems – further research is needed – but what this study emphasises is that the metalinguistic skill of phonological awareness helps to develop *all* languages.

2.4.2: Correlations Between L1 and L2 Knowledge and Skills

Having reviewed a number of more contemporary studies that attest to correlations between one aspect of CALP – namely, phonological awareness – across languages, this section will now examine correlations between other, more varied aspects. A longitudinal investigation by Van Gelderen et al. (2007), for example, explored the relationship between the L1 and L2 reading comprehension skills of 389 Dutch-speaking English students from Grade 8 to Grade 10. Tests were administered once in the middle of each school year, and results revealed strong positive correlations between L1 and L2 scores that began at r=0.85 in Grade 8, and increased over the

years. Since the participants had only been learning English for 1.5 years when the experiment started, Van Gelderen et al. (2007) interpreted this as undermining the threshold hypothesis by Cummins (1979a), and indicating instead that L2 learners can transfer L1 reading skills, such as metalinguistic knowledge of strategies and genre characteristics, despite their L2 level.

Of course, Cummins would argue that such knowledge constitutes CALP, rather than BICS, and thus that an individual who possesses it has already surpassed the lower threshold of bilingual competence in at least one of their languages. Once this has happened, the interdependence of cognitive/academic proficiency begins, and so the findings of this study arguably adhere to the threshold hypothesis, as well as the linguistic interdependence hypothesis. In research by Gebauer et al. (2013), the outcomes of Van Gelderen et al. (2007) were then reiterated when the L1 and L2 reading comprehension measures of 220 German-speaking elementary school students of English were found to correlate moderately in Grade 3, at r=0.47, and then more strongly in Grade 4, at r=0.56. In addition, L1 and L2 silent reading fluency was also observed to correlate at r=0.52 in Grade 4, and other studies (Domínguez de Ramírez & Shapiro, 2007; Geva et al., 1997; Ramirez, 2001) have reported similar findings for oral reading fluency.

That these studies demonstrate correlations between aspects of CALP other than phonological awareness is clear – even if the reasons behind these correlations are not – and Melby-Lervag and Lervag (2011) scrutinised this further. After conducting a systematic review of all research from 1976 to 2009 that assessed the relationships between oral language skills, phonological proficiency or reading comprehension in L1s and L2s, they identified a total of 47 studies with 52 samples comprising 4,413 L2 learners. The correlations presented in each of these studies were combined into five separate meta-analyses, and the researchers calculated large positive meta-correlations of r=0.60 between L1 and L2 phonological awareness, and r=0.54 between L1 and L2 decoding skills – with the studies in the previous section included. While there were too few samples to identify the meta-correlation between L1 and L2 reading comprehension, L1 decoding correlated with the latter at r=0.24, and importantly, this association was stronger in younger learners – perhaps due to their greater reliance on bottom-up reading skills.

Looking at L1 oral proficiency, on the other hand, it was not found to correlate with L2 reading comprehension at all (r=0.04), and this may be because the former entails BICS, whereas the latter entails CALP. Moreover, Melby-Lervag and Lervag (2011) detected only a small positive

meta-correlation of *r*=0.16 between L1 and L2 oral proficiency, and as this was not mediated by the socioeconomic status of the students – and therefore the amount of decontextualised language to which they were exposed – it provides support for the claim that BICS are typically not transferable. Another meta-analysis by Preevo et al. (2016) similarly intimated this when it determined that L1 oral proficiency correlated only weakly with early L2 literacy, at *r*=0.21, and even less so with L2 reading proficiency, at *r*=0.12. This analysis synthesised a total of 86 studies from 1976 to 2013, and was robust in design, so overall there appears to be extensive evidence in support of the theoretical constructs and hypotheses put forth by Cummins.

2.4.3: Correlations Between Dissimilar Languages

So far, the correlational research outlined in the previous sections has provided empirical data on the interdependence of CALP across orthographically similar languages – that is, languages that make use of alphabetic scripts. Whether this interdependence can be expected to extend to orthographically dissimilar languages, however, is an important question, and a number of studies have indicated that it can – at least in so far as phonological awareness (Bialystok, Luk & Kwan, 2005; Bialystok, McBride-Chang & Luk, 2005; Keung & Ho, 2009; Wang et al., 2005; Wang, Park & Lee, 2006; Wang et al., 2009) and morphological awareness (Wang et al., 2009; Wang, Cheng & Chen, 2006) are concerned. In this research, literacy skills and word decoding skills were discovered not to correlate across dissimilar languages on one occasion each – and orthographic processing skills not to correlate on multiple occasions, as might reasonably be predicted – but associations in the other areas were observed repeatedly, and the languages investigated were English, Spanish, Hebrew, Mandarin Chinese and Cantonese Chinese.

Furthermore, a study by Chuang et al. (2011) proved particularly compelling when it analysed the L1 and L2 reading proficiency of 30,000 Taiwanese Grade 9 students of English, and found them to be strongly positively correlated. The data of the participants was randomly selected from a larger sample compiled over a six-year period, and linear regression modelling revealed that measures of L1 reading proficiency accounted for 62.8% of the variation in measures of L2 reading proficiency. Not only this, but it also established that other predictors, such as the gender of the students, the location of their schools in rural or urban districts, and the amount of time spent learning English, had only a small effect on their L2 reading scores. That is, while female and urban students did significantly outperform male and rural students, the disparity

in their raw scores was negligible, and Chuang et al. (2011) assert that L1 reading proficiency determines L2 reading proficiency in both structurally similar and dissimilar languages.

2.4.4: The Limitations of Correlational Research

The inference of the studies summarised thus far is that L1 CALP provides a foundation for L2 learning, but a key point to consider when interpreting research on the relationship between L1 and L2 proficiency is that correlation does not equal causation. In other words, although the presence of positive correlations between the two means that a causal connection cannot be ruled out, it also does not guarantee its presence. Individuals with well-developed L1 CALP may be capable of acquiring well-developed L2 CALP for reasons other than the L1 CALP itself, and maintaining the L1 alongside the L2 may not facilitate its acquisition. To ascertain for sure whether it does or does not, experimental research that directly compares the effects of using the L1 at the same time as the L2 with not doing this on L2 learning outcomes is needed, and fortunately, there is a vast body of research in this area – namely, that which investigates the language learning effects of bilingual education, and is the focus of the following section.

2.5: Evidence From Bilingual Education Programmes

According to Dicks and Genesee (2016), bilingual education can be defined as any programme of study wherein two or more languages are utilised as media of content instruction, and early research into this practice in the Canadian context yielded promising results. One of the most seminal of these studies was the St. Lambert experiment by Lambert and Tucker (1972) which evaluated the L1 proficiency, L2 proficiency, mathematics knowledge, science knowledge, and general intelligence of two classes of English-speaking students of L2 French over a period of six years. In kindergarten, these students were exposed entirely to L2 instruction – with classes in the English language arts constituting the only exception – but their L1 instruction increased every year, until the balance became equal by Grade 5. In terms of findings, comparisons with control groups in monolingual schools then showed the L1 proficiency, subject knowledge and IQ of the experimental group to be equivalent, while their L2 proficiency was superior.

What this means is that utilising the L1 in conjunction with the L2 had no detrimental effects for these students, yet benefitted their L2 skills, and Lambert et al. (1973) also observed that their progress in science, maths and intelligence was sometimes greater. It is worth noting at

this point that participants were not randomly assigned to the experimental condition, as a result of being volunteered by their parents, and that they were of high socioeconomic status, and encouraged to pursue bilingualism. As such, the generalisability of their outcomes to less motivated or affluent learners cannot be assumed, but the tentative success of the bilingual model in Canada – combined with claims that all children should be educated in their mother tongue (UNESCO, 1953) – contributed to its implementation in the United States (Chalmers & Murphy, 2022). In 1974, amendments to the Bilingual Education Act of 1968 mandated that all states must provide bilingual education to language minority students who desired it, and although there was some opposition to this, it enabled widespread empirical research.

In fact, Cohen (1974) reported on a partial replication of the St. Lambert experiment wherein English-speaking children were placed in Spanish immersion kindergarten and Grade 1 classes, and assessed for their L1 proficiency, L2 proficiency and maths skills. This study was conducted over a shorter period of two years, and the participants were taught only in their L2, but their L1 was arguably sustained by their home and societal environments, and analysis established that their maths and English language skills were on a par with peers in L1-only classes, while their Spanish language skills were on a par with peers in L2-only classes. Essentially, L1 and L2 use seemed to hasten their L2 learning – whilst not reducing their L1 or content learning – and a collection by Krashen and McField (2005) of five meta-analyses into the efficacy of bilingual education (Greene, 1997; McField, 2002; Rolstad et al., 2005; Slavin & Cheung, 2005; Willig, 1985) also found it to be linked to equal or enhanced academic and linguistic outcomes.

Admittedly, these meta-analyses did not distinguish between types of bilingual programmes, and simply compared them as a whole to all-English instruction, but their average effect sizes were nonetheless similar, with a range of d=0.18 to 0.33. A more recent systematic review of bilingual education in Europe by Reljić et al. (2015) then corroborated this when it employed a random-effects model to synthesise the findings of five of its seven studies, and established a small positive effect of bilingual education over L2-only education, at g=0.23. This effect was on the reading-related skills of language minority children, and the L1s appraised were Urdu, Asturian, Basque, Catalan and Turks, alongside the L2s of English, Spanish and Dutch. For Reljić et al. (2015), their review backs bilingual education when it uses the L1 of minority students, and as research in the North American context also showed it to be successful when using the L2 of majority students, it seems to work well with both kinds of potential participant.

2.5.1: Evidence From Randomised Studies

Of course, a primary concern when analysing research on the effects of bilingual education in comparison to monolingual education is that the participants – or their parents – have chosen to enrol in one type of programme or the other. Systematic differences in their characteristics across conditions are thus likely, and although they can be matched for some known variables, such as language learning motivation, socioeconomic status or non-verbal IQ, it is impossible to match them for unknown or unassessable variables. The only way to do this with certainty would be to allocate them randomly to bilingual or monolingual learning conditions, but doing this against their preferences would naturally pose ethical problems, and hardly ever happens organically. One exception to this is the large-scale longitudinal research conducted by Steele et al. (2017) in Portland, USA, where bilingual programmes are so oversubscribed that the city uses a lottery system to assign applicants either to them, or to English-only alternatives.

In this study, the researchers report on 1,946 students, 864 of whom were placed in one-way or two-way dual language programmes, and 1,082 of whom were placed in unilingual English programmes. In the one-way option, the students were all native speakers of English, and in the two-way option, they were native speakers of either English, or the partner languages of Spanish, Russian, Japanese and Mandarin. Data were collected over multiple years, and it was discovered that the students in the bilingual programmes outperformed their peers in Grade 5 and Grade 8 English reading skills by the equivalent of first seven, and then nine, months of learning. They also attained intermediate proficiency in their L2s by Grade 8, which their peers did not in Spanish, and these results reiterated those of another randomised study by Barnett et al. (2007). All in all, this methodologically sound research supports the less rigorous studies of the previous section, and indicates that bilingual education – that is, cultivating the L1 along with the L2 – advantages the L2 learning of both minority and majority language users.

2.5.2: What Does This Mean for L2 Teaching?

Altogether, research into bilingual education programmes has revealed that the individuals in them perform as well as, or better than, those in non-bilingual programmes, at least in terms of their L1 development, L2 development, academic achievement and IQ. This suggests that there is a causal relationship between attending such a programme, and improved academic

or linguistic outcomes, and the implication of this – as Cummins (1998) proposed – is that L2 classrooms in non-bilingual contexts might benefit from embracing the L1s of their learners, and abandoning the monolingual principle. Precisely how they ought to do this is a question that requires further research – not least because bilingual education uses the L1 as a medium of instruction, rather than a pedagogical tool, which may be less effective in L2 learning – and one way that Cummins (2001) recommends is examining how L1s are used instinctively by L2 teachers and learners. In other words, the functions to which they are automatically put, even when L2-only policies prohibit this, and some of these functions will be discussed below.

2.6: How Multilingual Learners Use Their L1s

Numerous studies (Anton & DiCamilla, 1998; Brooks & Donato, 1994; de la Campa & Nassaji, 2009; Kobayashi, 2003; Storch & Wigglesworth, 2003) have offered evidence of the L1 serving as a scaffold for L2 task achievement in both foreign language and immersion classrooms. In a paper by Swain and Lapkin (2000), for example, 22 pairs of Grade 8 English-speaking French immersion students were asked to complete two L2 writing tasks, while using their L1 however they may like, and analysis of their conversations identified its functions as moving the task along, focusing on L2 grammar or lexis, and facilitating interpersonal interaction. A later study by Scott and de la Fuente (2008) then expanded on the latter of these when it discovered that pairs of students who were permitted to use their L1 while writing L2 grammar explanations collaborated more effectively than those who were not, and for the researchers, this implied that L1 use alleviates cognitive resources, and helps learners to improve their L2 output.

De la Colina and del Pilar Garcia Mayo (2009) asserted similarly when they found that twelve pairs of Spanish-speaking English students used their L1 to engage in metatalk about the task or its language when producing L2 texts, but crucially, only one of these studies assessed the relationship between L1 use and L2 quality. When doing this, Swain and Lapkin (2000) noticed that pairs who created above median texts in terms of both language and content took fewer conversational turns in their L1 than pairs who created below median texts, and this discovery was echoed by Van Weijen et al. (2009). In this research, the frequency and function of L1 use by twenty Dutch-speaking students writing four L2 essays was compared to their quality on a 100-point scale, and not only did only one function relate to this quality – that of using the L1 for meta-commenting – but it did so negatively, with a regression weight of β=-.50. What this

implies is that L1 use may not be advantageous, and so while analysing how it naturally occurs may allude to its pedagogical value in non-bilingual classrooms, this cannot be assumed.

2.7: Recommendations for Pedagogical L1 Use

Despite the conclusion of the previous section, and the inference that the effects of L1 use on a variety of L2 learning outcomes need to be established before it can be advocated, there is nonetheless a substantial body of literature that advises teachers on how to incorporate the L1s of their learners into L2 lessons. Much of this literature follows from the multilingual turn in language teaching (Conteh & Meier, 2014; May, 2014), which constitutes a move away from the monolingual ideologies of the past, and toward an appreciation of multilingualism as the global norm. Cummins and Early (2011), for instance, put forth the concept of 'identity texts', which are dual-language creative works that multilingual individuals can construct in written, spoken, signed, visual, musical, dramatic or multimodal forms, and which can function as an identity-affirming source of pride for them. Other theorists (Celic & Seltzer, 2012; Espinosa et al., 2016; Hesson et al., 2014) have posited L1-mediated teaching methods that centre around notions of 'translanguaging', and this will be further explored in the subsequent section.

2.7.1. Pedagogical Translanguaging

Originally known as 'trawsieithu', the term translanguaging was developed by Williams (1994) in the context of Welsh bilingual education, and used to describe the pedagogical practice of switching between languages for instructional input and output. Following this, the term was extended by academics in the United States (Garcia, 2009; Garcia and Wei, 2014; Paulsrud et al., 2017; Wei, 2018) to reflect the interlingual communicative acts of bilingual speakers, and they were said to possess dynamic, rather than additive or dominant, bilingualism. What this means is that instead of their linguistic abilities being perceived as equivalent to those of two monolinguals (Grosjean, 1989), they were thought to signify a 'multicompetence' (Cook, 1992; 1995) that derives from a single linguistic repertoire. In the strong version of translanguaging, this repertoire is comprised of linguistic features that individuals utilise selectively, but which do not belong to named languages per se (Otheguy et al., 2015), and in the weak version, its advocates acknowledge the social reality of languages, but call to relax their boundaries.

The intention of both of these versions is for bilingual students to be able to employ their full

linguistic repertoire in any situation, and for the separation of languages in education – which Cummins (2008) refers to as the 'two solitudes assumption' – to not be so rigidly enforced. In complementary schools in the United Kingdom, Creese and Blackledge (2010) showed this to be a naturally-occurring educational approach, which they personally advocate, and from this, the practice of pedagogical translanguaging has evolved. As Cenoz and Gorter (2021) explain, this involves the planned use of instructional strategies, in either language or content learning classrooms, that appeal to the full linguistic repertoire of multilingual learners in an inclusive 'translanguaging space'. Priority is placed on the promotion of metalinguistic awareness, and in order to ensure that majority languages do not dominate minority ones, a balance needs to be achieved in class, and minority language-only 'breathing spaces' created outside class.

2.7.2: The Limitations of These Practices

Overall, pedagogical translanguaging supports extensive use of the L1 in L2 classrooms, much like other practices born from the multilingual turn (see Chumak-Horbatsch, 2012). In keeping with this, research into the effects of L1 use on student well-being has suggested that it helps them to feel appreciated for their multilingualism (Cummins et al., 2015; Dakin, 2012; Kenner et al., 2008; Parke et al., 2002), but importantly, this says nothing of its effects on academic or linguistic outcomes. Bilingual education may be associated with greater achievement in these areas, but any number of factors could account for that, beyond concurrent L1 and L2 use. If educators are to trust, therefore, that the implementation of multilingual pedagogies – which requires both time and training – will be beneficial for L2 learners, then research that directly compares them to monolingual alternatives must be conducted. This is the key to identifying whether using the full linguistic repertoire of multilingual students is actually the best means by which to aid their academic and linguistic learning, but studies that do this are lacking.

2.8: Experimental Research on the Effects of L1 Use

To demonstrate precisely how lacking, a systematic review by Chalmers (2019) that compiled experimental and quasi-experimental research into the effects of L1-mediated teaching in L2 settings on L2 proficiency and/or academic attainment located only ten studies. This is despite the fact that these studies were not required to evaluate L1-mediated teaching approaches in comparison to L2-mediated teaching approaches, but rather only L1-mediated approaches on

their own, and the ten studies examined a range of teaching strategies and outcomes. Three of them, for example, explored the effects of L1 use on measures of L2 reading, and produced conflicting results (Walters & Gunderson, 1985; Sanchez, 2004; Huennekens, 2013). Another (Yiakoumetti, 2006) found in favour of L1 use when contrastive L1 and L2 analysis was shown to enrich spoken and written L2 accuracy, and Chalmers (2014) concluded differently when he discovered that discussing a story in the L1 or the L2 did not affect its L2 written retelling.

2.8.1: Using the L1 to Teach L2 Vocabulary

The five other studies in the systematic review (Camó & Ballester, 2015; Lee & Macaro, 2013; Lugo-Neris et al., 2010; Sieh, 2008; Tonzar et al., 2009) were more uniform in both their focus and their findings, with each of them investigating L1-mediated approaches to L2 vocabulary teaching. Although only one of them adopted a randomised design, and the scale of three was relatively small, the consistency of their support for L1 use in a range of contexts (see Table 1) suggests that it might facilitate the teaching of L2 lexis. Naturally, these five studies alone are insufficient to conclude this with any certainty, and much more evidence has to be gathered, but they provide what Chalmers and Murphy (2022) call a "promising avenue for exploration" (81), and constitute the basis for this systematic review. The one by Chalmers (2019) targeted research from 1980 to 2015 – specifically in L2 primary and pre-primary school settings – and this review intends to build on that by both expanding and refining its eligibility criteria.

Table 1: Summary of the Five Studies on L2 Vocabulary Teaching

Study	Design	Context	Outcome Measures	Subjects	Finding
Camó and Ballester, 2015	Non-randomised comparison	Primary school EFL programme	Listening: Receptive vocabulary	34 Catalan L1 speakers, aged 10-11	Favoured L1 use
Lee and Macaro, 2013	Non-randomised comparison	Primary school EFL programme	Reading and writing: Receptive vocabulary	443 Korean L1 speakers, aged 12	Favoured L1 use
Lugo-Neris et al., 2010	Randomised crossover trial	Summer migrant ESL programme	Listening, speaking and writing: Productive and receptive vocabulary	22 Spanish L1 speakers, aged 4-6	Favoured L1 use

Sieh, 2008	Non-randomised comparison	Primary school EFL programme	Listening: Receptive vocabulary	64 Mandarin or Taiwanese L1 speakers, aged 9-10	Favoured L1 use
Tonzar et al., 2009	Non-randomised comparison	Primary school EFL and GFL programmes	Writing: Productive vocabulary	123 Italian L1 speakers, aged 9-10	Did not favour L1 use

Key: EFL=English as a Foreign Language; ESL=English as a Second Language; GFL=German as a Foreign Language

Chapter Three: Methodology

3.1: The Research Question

The purpose of a systematic review, as stated by Macaro (2019), is to comprehensively locate, assess and synthesise all empirical research on a specific topic, so that theoretical claims – and in this case, explicit pedagogical recommendations – can be evaluated in light of evidence. For Dickson et al. (2017), they differ from more traditional literature reviews in that their methods are more transparent, their search strategy more exhaustive, and their attempts to minimise reviewer bias more pronounced. All of their included studies, for example, are judged for their reliability before unifying their findings, and multiple reviewers are employed to corroborate subjective interpretations where possible. In this review, every effort has been made to follow these requirements, with the author adhering to the PRISMA reporting guidelines by Page et al. (2020) as far as possible (see Appendix 1), and its aim is to ascertain the nature and extent of empirical research, beyond that already located, in relation to the following question:

- What are the effects of using the L1 in comparison to the L2 to teach target language vocabulary items in primary or secondary school settings?

The hope is that this research, and the inferences drawn from synthesising its outcomes, will help L2 teachers and policy-makers to make evidence-based decisions about the appropriacy of L1-mediated approaches to L2 vocabulary teaching. In this chapter, the methodology of the review will be outlined in detail, starting with an overview of its eligibility criteria, sources of information, search strategy and study selection process, and then proceeding to a discussion of its data extraction methods, risk of bias assessment, and approach to data synthesis.

3.2: Eligibility Criteria

Table 2 presents the inclusion/exclusion criteria that were determined for this review.

Table 2: The Inclusion and Exclusion Criteria

Criteria	Rationale
Bibliographic Information **Include 1:** Sufficient reference. **Exclude 1:** Insufficient reference.	Without sufficient bibliographic information, such as the complete title, list of authors or publication date, locating a study is likely to be time-consuming and/or futile.
Focus of the Systematic Review **Include 2:** Relevant to assessing the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition. **Exclude 2:** Not relevant to this.	This review seeks to contrast the effects of using either the L1 or the L2 to teach L2 vocabulary items, and therefore studies that do not directly compare the two are irrelevant.
Publication Date **Include 3:** Published between 1980 and 2023. **Exclude 3:** Published before 1980.	The linguistic interdependence hypothesis, which speaks to the role that the L1 may play in L2 acquisition, was posited by Cummins in 1979. His distinction between BICS and CALP, which lent nuance to this hypothesis, then followed in 1980 – along with his notion of a CUP – and so research that responds to these ideas will be from after that time. In addition, Chalmers (2019) examined the effects of using L1-mediated teaching approaches in L2 primary and pre-primary school settings, and focused on the years 1980 to 2015. As this review aims to build on his findings by exploring the effects of such approaches on L2 vocabulary acquisition in particular, a similar starting date with an expanded range seems appropriate.
Study Design **Include 4:** Intervention studies in the form of randomised controlled trials or quasi-experiments. **Exclude 4:** Non-intervention studies.	Since this review aims to determine whether teachers should be advised to use L1-mediated or L2-mediated approaches to L2 vocabulary teaching, it is necessary to examine empirical evidence of their effects.

Study Setting **Include 5:** Studies of formal education, focusing on typical school/student populations. **Exclude 5:** Studies of informal education, such as private language centres or complementary schools, and studies that focus on non-typical school/student populations, such as those with special educational needs.	Students with special educational needs are likely to have requirements that set them apart from mainstream learners, and may confound research into the effects of L1-mediated and L2-mediated teaching approaches on L2 vocabulary acquisition for learners without special needs by producing quite different results. Note: Studies that are conducted in contexts where students with special educational needs may be included in otherwise typical populations, such as mainstream primary or secondary schools, will not be excluded.
Study Participants **Include 6:** Conducted with primary and/or secondary school-aged children. **Exclude 6:** Conducted with pre-primary school-aged children, university students, or adults.	The systematic review by Chalmers (2019) was officially a scoping review, designed to examine the nature and extent of empirical research into the effects of L1-mediated teaching approaches in L2 educational settings in general. As it found some evidence to suggest that using the L1 to teach L2 vocabulary is advantageous at the primary level, however, it is sensible to ascertain whether (a) additional research has been undertaken on this topic, and (b) this finding holds true at the secondary level. Note: Given that the age of students at different levels of education varies slightly between – and sometimes within – countries, the definition of primary or secondary school-aged students in the context of each study will determine its inclusion, rather than their chronological age.
Study Outcomes **Include 7:** The outcomes are quantitative measures of L2 vocabulary acquisition, either in the form of target L2 word knowledge, or the comprehension of texts that contain target L2 words. **Exclude 7:** The outcomes are qualitative measures of L2 vocabulary acquisition, or quantitative measures	Although qualitative measures of L2 vocabulary acquisition could help to shed light on how students feel about L1-mediated and L2-mediated vocabulary teaching approaches – and therefore whether they would be well-received in schools – the focus of this review is on whether L1-mediated or L2-mediated approaches facilitate greater vocabulary acquisition.

that are either entirely self-reported, or assess alternative outcomes.	
Publication Type **Include 8:** Any type of publication, including grey literature, that reports directly on empirical research. **Exclude 8:** Any type of publication that does not do this, such as book reviews or editorials.	This review aims to synthesise all research on the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition over a 43-year period, regardless of the form that it takes.

An additional consideration that was ultimately not employed as an eligibility criterion is that of the language used for the publication of empirical research. As Moher et al. (2003) explain, limiting a systematic review to studies reported only in English can serve to arbitrarily exclude relevant research, after all, and to bias its findings by rendering them non-comprehensive. On this occasion, while the search terms selected for the review were English-only for pragmatic reasons (see Section 3.4), any abstracts identified by them in non-English languages were still considered for inclusion. To enable this, a general understanding of their content was gained via Google Translate, and this proved sufficient for making inclusion/exclusion decisions.

3.3. Information Sources

Before conducting this review, online databases of existing or proposed systematic reviews in the field of education were searched, so as to ensure that it would be contributing something new (Petticrew & Roberts, 2006). The databases examined were the EPPI-Centre Database of Education Research (EPPI-Centre, 2019), the Campbell Collaboration Online Library (Campbell Collaboration, 2023), and the International Database of Education Systematic Reviews (IDESR, n.d.). The keywords used in this search were the generic 'vocabulary' or 'vocabulary teaching', due to more specific terms yielding zero findings, and once the originality of the review had been established, a protocol document (see Appendix 2) was created. In this, its main sources of information were identified as five bibliographic databases, accessed electronically through the Bodleian Library, that cover multiple linguistics collections, and these databases are:

- ProQuest Social Science Premium Collection
- British Education Index
- Web of Science
- PsycINFO

- Scopus

In addition to the electronic searching of online bibliographic databases, Brunton et al. (2017) also advise that other sources of information be utilised, such as the reference lists of already located studies. This can help to mitigate the negative effect of errors or inconsistencies in the indexing of articles in bibliographic databases (Siddaway et al., 2019) – as well as expose those that have not been included – and in this review, backwards citation searching was employed by consulting the reference lists of all research selected for inclusion, while forwards citation searching was employed using the 'cited by' function on Google Scholar. Given limitations of time, hand searching journals was considered impractical, but studies known to the reviewer, such as those identified by Chalmers (2019), were assessed against the eligibility criteria.

3.4: The Search Strategy

According to Petticrew and Roberts (2006), it is crucial to strike a balance between sensitivity and specificity when developing the search strategy for a systematic review, so that as many relevant studies are retrieved as possible, and as few irrelevant ones. One way of doing this is to carry out a series of scoping searches that pilot a range of potential terms, and ensure that suitable studies are being detected (Dundar & Fleeman, 2017a). In this review, this was done until three fields of terms (see Appendix 3) were devised in relation to the concepts of L1s, L2s and vocabulary, and 120 were selected in total. Although this may seem excessive, extensive piloting of their associated Boolean strings (see Appendix 4) found the lower-frequency terms to identify only a small number of additional studies, and so they were judged to deepen the sensitivity of the search without jeopardising its specificity, and to be worthy of inclusion.

Naturally, the Boolean strings needed to adopt the field codes of the bibliographic databases for which they were designed, but in general, the three fields were united by 'AND' functions, and their individual terms by 'OR' functions. Non-standard expressions were included in both hyphenated and unhyphenated form, and the truncation symbol * was used to accommodate any morphological variation in the terms. Two research consultants from the Department of Education attested to the appropriacy of this approach, and the five databases were searched on 5th March 2023. In total, 5,708 records were returned, and these were first downloaded as .RIS files, and then uploaded to Rayyan, which is an online platform for the administration of systematic reviews (Ouzzani et al., 2016), and screened for either inclusion or exclusion.

3.5: The Selection Process

When screening the often substantial amount of records uncovered by a systematic review, it is typical for a two-stage process to be followed (Dundar & Fleeman, 2017b), and for detailed accounts to be kept of the studies excluded at each of these stages, as is outlined below.

3.5.1: Stage 1: Title/Abstract Screening

A useful feature of the Rayyan software is that it highlights possible duplicates of records that have been sourced from more than one database, and so the preliminary step of stage 1 was to compare these carefully, and delete as necessary. After this, the titles and abstracts of the remaining records were screened against the eligibility criteria, and those that failed to meet at least one of them were labelled as 'excluded', with the primary criterion that supported this decision noted. Any records that could not be excluded based solely on this information were marked as 'included', and in order to reduce the possibility of reviewer subjectivity, fatigue or error biasing the selection process (Brunton et al., 2017), 10% of the 3,503 records remaining after de-duplication were blindly double-screened. This was done by a Delta-qualified English language teacher who was briefed on the purpose of the review, its eligibility criteria, and how to use Rayyan, and inter-rater reliability was high, at 99.4%. The two conflicting decisions (see Appendix 5) were then resolved through discussion, with both studies finally excluded.

3.5.2: Stage 2: Full-Text Screening

During stage 2 of the screening process, the bibliographic information of each study included in stage 1 was used to locate its full report, either through the Bodleian Library online system, inter-library scan requests, or the academic website ResearchGate.net. On one occasion, the personal webpage of the author was used, and these reports were then evaluated against the eligibility criteria of the review, with a spreadsheet developed to document their inclusion or exclusion outcome (see Appendix 6). Once again, the second reviewer assessed 10% of these reports, and since inter-rater reliability was 100%, all of those remaining – including the ones located through forwards or backwards citation – were screened by only the main author.

3.6: The Data Extraction Process

After the total number of records to be included in the review had been determined, each of them was read in detail, and key information extracted. This was done by means of a bespoke data extraction form (see Appendix 7) that was originally developed with reference both to the example provided by Petticrew and Roberts (2006), and that utilised by Chalmers (2019) in his systematic review. Multiple versions of it were piloted with a small number of randomly selected studies in order to ensure that it was appropriately transparent, comprehensive and flexible – as is suggested by Fleeman and Dundar (2017) – and the final form comprised four sections, and was used to collect the descriptive and analytical data outlined in Table 3.

Table 3: The Data Extracted from Included Studies

Section	Type of Information	Subsections
A	Administrative	In-text Reference, Study Title, Identification Method, Publication Type, Publication Language
B	Contextual	Study Location, Educational Setting, Type of Language Programme, Study Participants
C	Intervention	Study Design, Research Question(s), Number of Conditions, Names and Sizes of Each Condition, Method of Participant Allocation to Each Condition, L1-mediated Vocabulary Teaching Approach(es), L2-mediated Vocabulary Teaching Approach(es), Administrator of Each Condition, Target L2 Vocabulary, Study Duration
D	Outcome	Outcome Measure(s), Pre-Test Measure(s), Quantitative Data for Each Measure In Each Condition, Effect Size(s), Study Finding(s)

3.7: Risk of Bias Assessment

Since this review constitutes a 'best-evidence synthesis' (Slavin, 1986), in which all studies that adhere to minimal standards of adequacy and relevance are included, it is essential that their methodological quality also be assessed. Failing to do so would risk conflating strong evidence with weak – thereby leading to inaccurate or misleading conclusions (Gorard, 2014) – and so after reviewing the six risk of bias assessment tools recommended for use with quantitative studies by Petticrew and Roberts (2006), the Quality Assessment Tool for Quantitative Studies by the Effective Public Health Practice Project (see Thomas et al., 2004) was chosen. Hereafter

referred to as the EPHPP tool, it was originally applied to public health research, but as it was found to be well suited to the evaluation of randomised controlled trials (Armijo-Olivo et al., 2012), which are commonplace in education, it was deemed appropriate for this review.

When using this tool, assessors are required to judge the methodological quality of a study as strong, moderate or weak in eight separate areas, but since the final two do not contribute to the global rating of the study, only the first six will be utilised here. Minor adjustments to the language used exchanged the term 'controlled clinical trial' in the Study Design section of the form with 'non-randomised comparison', because schools are non-clinical locations, and the question 'Were there important differences between groups prior to the intervention?' in the Confounders section was extended on the advice of Chalmers (2019) to add 'that were known to the researchers'. This is because unknown differences cannot be controlled, and ultimately the six sections (see Appendix 8) gathered information on the features shown in Table 4.

Table 4: Summary of the EPHPP Quality Assessment Criteria

Section	Area of Assessment	Focus of the Assessment
A	Selection Bias	Whether the individuals selected to participate were representative of the target population, and how many of them agreed to participate.
B	Study Design	The design of the study, and whether it controlled for potential biases through randomisation
C	Confounders	Whether there were important differences between the groups of participants, and whether these confounders were controlled
D	Blinding	Whether the outcome assessors were aware of participant group allocation, and whether the participants knew the research question(s)
E	Data Collection Methods	Whether the data collection tools were shown by the researcher(s) to be valid and/or reliable
F	Withdrawals and Drop-outs	Whether participant withdrawal was reported, and how many of them completed the research

Using the modified EPHPP tool, risk of bias assessments were carried out for all of the studies as soon as their data had been collected, and as with the screening process, 10% of them were assessed by a second rater. Initially at 97.1%, inter-rater reliability on this reached 100% after discussion, and the remaining studies were then assessed by the main researcher only.

3.8: Synthesis of Results

Prior to performing this systematic review, the work of Chalmers (2019) had already shed light on the type of research conducted into L1-mediated approaches to L2 vocabulary teaching at the primary level. Less was known about the nature and extent of research at the secondary level, but had it proven to be sufficiently similar, then effect sizes would have been combined in a meta-analysis (Brown & Richardson, 2017). Since this was not the situation – with studies employing diverse interventions, comparators and outcome measures – a narrative approach to synthesis was instead adopted, and findings presented according to their principal themes (Thomas et al., 2017). What this means, as Petticrew and Roberts (2006) clarify, is that studies were first organised into logical categories, then examined within those categories, and finally across categories, in a three-step process the results of which are in Chapters 4 and 5.

Chapter Four: Results

4.1: Study Selection

As mentioned previously, electronic searches of the five bibliographic databases identified for this review produced a total of 5,708 records, while an additional five were already known to the reviewer from Chalmers (2019). Following their de-duplication, first stage screening of the outstanding 3,503 records then led to 3,409 being excluded for the reasons detailed in Figure 5, which depicts the study identification process in flow diagram form (Page et al., 2020), and of the 94 records that qualified for second stage screening, 87 were retrieved in their full-text format. At this point, 57 more were excluded – largely on account of being with tertiary-level adult participants – and an extra 15 were located through forwards (n=9) and backwards (n=6) citation. Altogether, this amounted to 45 studies that fulfilled all eight of the eligibility criteria (see Appendix 9), and their descriptive data will be summarised in the following sections.

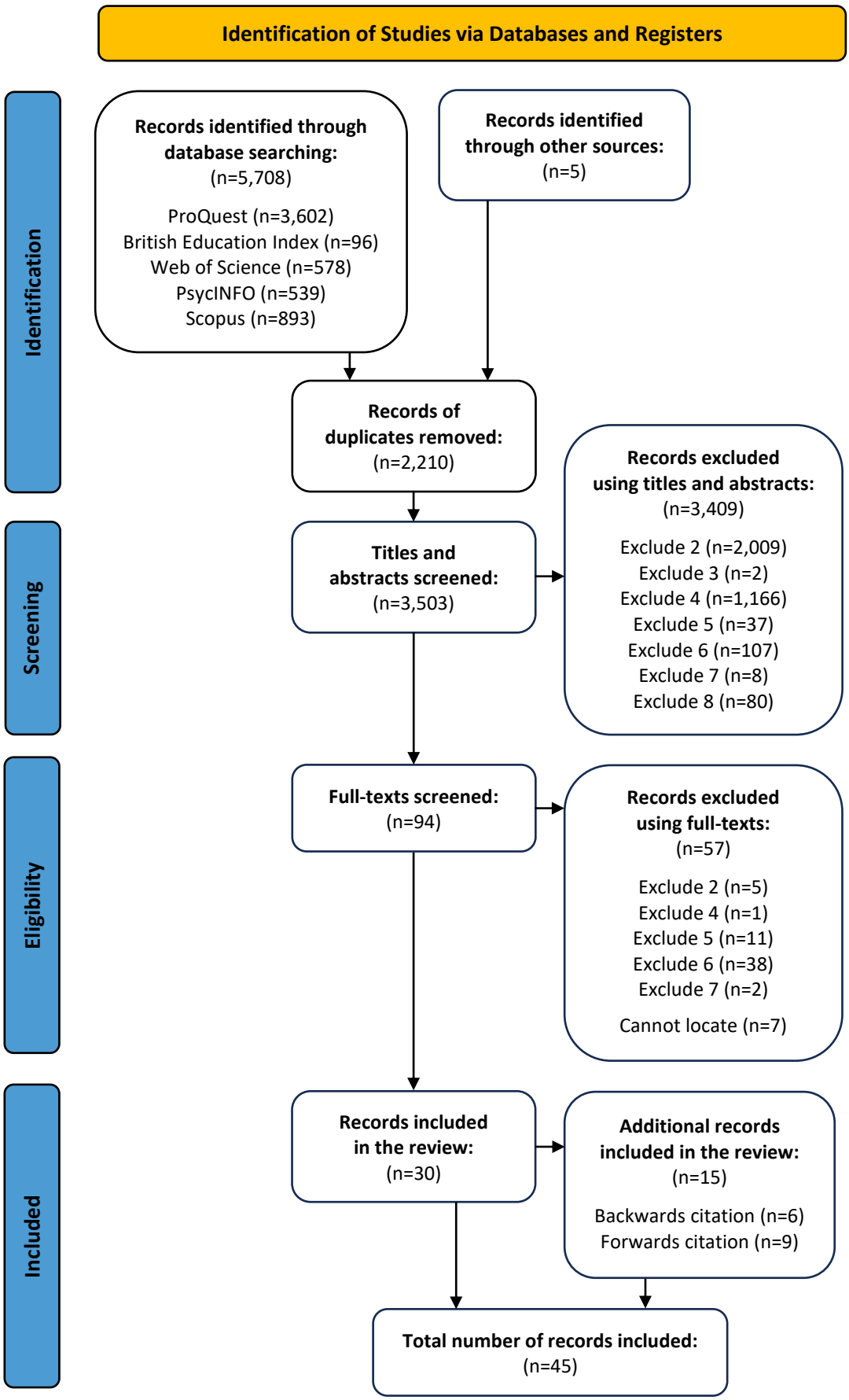


Figure 5: The PRISMA Flow Diagram

4.2: Study Characteristics

To begin, it is worth acknowledging that four of these 45 studies reported on more than one experiment, so that 51 were described in total. In some cases, all of those within a single study were considered germane (two in Study 36 (S36) and four in S42), but in others (S31 and S35), only one of them was. This was due to S31 running one of its experiments with adults – while S35 focused one on grammar acquisition – and so overall, 49 experiments were relevant.

4.2.1: Publication Date and Type

With regard to the publication dates of the included studies, none of them were from before 1997, despite the seminal work of Cummins substantially predating that. As Figure 6 indicates, research into the effects of L1-mediated vocabulary teaching in comparison to L2-mediated vocabulary teaching on L2 vocabulary acquisition has increased steadily since the early 2000s, but it is still relatively scarce, and – as section 4.2.3 will demonstrate – diverse in nature.

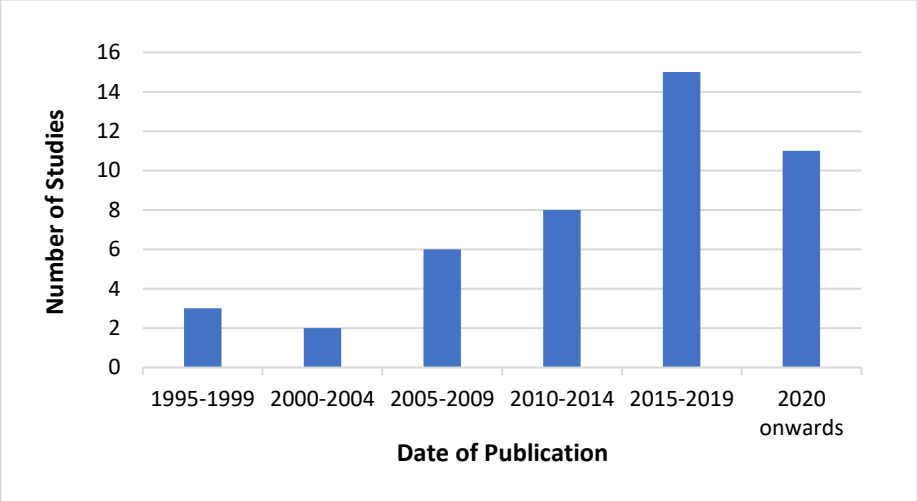


Figure 6: The Included Studies by Publication Year

The publication type of the 45 studies was mainly articles in peer-reviewed journals (see Table 5), but a considerable number were Master's dissertations or doctoral theses as well.

Table 5: The Included Studies by Publication Type

Publication Type	Quantity
Journal article	38
Master's dissertation	3
Doctoral thesis	3
Book chapter	1

4.2.2: Language Programme Type and Location

Turning to the type of language programme that the participants in the studies were exposed to, the vast majority studied English as a Foreign Language (EFL) in mainstream primary (n=14) or secondary (n=31) schools around the world. On one occasion, they also learned English as an Additional Language (S14), and either German in addition to English (S42), Spanish (S16) or Italian (S25) as foreign languages. French was then acquired as both a foreign language (S22) and a second language (S8) in the varied educational contexts of the UK and Canada, and to provide further information on the range of countries involved in the research, Table 6 shows that Spain, China and Taiwan were best represented, with Saudi Arabia, Germany and Israel not far behind. After that, all other countries were surveyed only once or twice, resulting in a total of 24, and more studies (n=9) concentrated on urban areas than rural ones (n=1).

Table 6: The Included Studies by Country

Country	Multiple Studies	Country	Single Studies
Spain	5	Mexico	1
Taiwan	4	Catalonia	1
China	4	Canada	1
Saudi Arabia	3	Poland	1
Germany	3	Sweden	1
Israel	3	Greece	1
Turkey	2	Thailand	1
The UK	2	The Netherlands	1
Korea	2	Malaysia	1
Iran	2	Italy	1
Japan	2	Austria and Norway	1
Belgium	2		

4.2.3: Study Design and Focus

In terms of the research designs employed by the included studies, they were divided equally between experimental and quasi-experimental approaches, with the former (n=23) qualifying as randomised controlled trials (RCTs), and the latter (n=22) as non-randomised comparisons (NRCs). Most of the RCTs (n=16) made use of cluster randomisation, which means that intact classes were randomly allocated to experimental or control conditions, but some (n=7) made use of individual randomisation, at the level of each participant. In addition, nearly all of them (n=21) were randomised parallel trials, where each group of subjects is exposed to only one of the study interventions, but two (S14 and S22) were randomised crossover trials, where all of the subjects experienced all of the conditions. This was the case for four of the NRCs as well (S10, S19, S26 and S40), and all 45 of the studies evaluated intervention effects by comparing pre-test, post-test or delayed post-test scores across conditions, as Table 7 illustrates.

Table 7: The Structure of the Included Studies

Study Structure	Study Design	
	RCTs	NRCs
Pre-test/post-test	6	6
Post-test/delayed post-test	6	5
Pre-test/post-test/delayed post-test	7	9
Post-test only	4	2

It should be noted that while many of the NRCs (n=16) compared the effects of interventions on intact classes, some (n=3) controlled for confounding variables by generating comparison groups on the basis of participant characteristics, and others (n=3) did not say. Regarding the interaction between study design and study focus, the range of interventions investigated was considerable (see Table 8), and overall, those most frequently researched were the use of (1) L1/L2 audio or subtitles while watching television, (2) L1/L2 oral glossing while listening, and (3) L1/L2 written glossing while reading, and their effects on L2 vocabulary acquisition.

Table 8: The Focus of the Included Studies

Type of Intervention	Total	Study Design	
		RCTs	NRCs
The L1 translation method vs. the pictorial/video method	5 studies (3 primary and 2 secondary)	2	3
Watching television with various combinations of L1/L2 audio, L1/L2 subtitles or oral definitions	15 studies (5 primary and 10 secondary)	9	6
Listening with various kinds of L1/L2 glossing, storytelling or contrastive analysis	8 studies (2 primary and 6 secondary)	5	3
Reading with various kinds of L1/L2 glossing or writing activities	13 studies (2 primary and 11 secondary)	4	9
Reading with content discussion, translation or vocabulary tasks	2 studies (2 secondary)	1	1
Translation-based learning activities vs. textbook-based learning activities	2 studies (2 primary)	2	0

Note: All of these interventions were delivered by either the researcher(s), or the regular language teacher(s) of the participants

4.2.4: Study Size and Duration

The number of participants involved in the included studies ranged from 6 in S20, which was intended as a precursor to a larger experiment not yet conducted, to 443 in S31. Table 9 shows that the majority of them (n=31) had fewer than 100 subjects in total – thereby limiting their ability to make causal claims (Gorard, 2014) – and importantly, the interventions were always delivered to the participants in groups. These interventions ranged in length from as short as two minutes and 40 seconds in S23, over four separate 40-second treatments, to as long as 8 hours and 35 minutes in both S35 and S38 – which adopted a similar procedure by two of the same researchers – over twenty-four 22-minute treatments. When pre-testing, post-testing, and delayed post-testing were considered, however, then these lengths increased (see Table 10) to a range from around one hour (S9 and S25) to a full academic year (S35 and S38).

Table 9: The Size of the Included Studies

Number of Participants	Number of Studies
0-50	12
51-100	19
101-150	8
151-200	3
201-250	2
251-300	1
301-350	0
351-400	0
401-450	1

Table 10: The Duration of the Included Studies

Duration	Number of Studies
Less than 1 month	23
1-2 months	11
2-3 months	4
3-4 months	1
4-5 months	1
5-6 months	1
6+ months	2
Unclear	2

4.2.5: Participant Proficiency, Gender and L1 Background

Surprisingly, the L2 proficiency of the participants was often unreported (n=16), or described using vague language like 'low proficiency' (n=15) and context-specific language like 'Level C' (n=2). On one occasion (S21), proficiency was self-assessed, but in the remaining studies it was helpfully defined using either the Common European Framework of Reference (CEFR) (n=10), or the five-point scale of the Department for Education in the UK (n=1). In all cases, it ranged essentially from beginner to intermediate level, and the only exception was S14, where 37.5% of the participants were regarded by their teachers as already being 'fluent' in English.

Concerning gender, this was typically either unconfirmed (n=22) or mixed (n=17), although six studies did have exclusively male (n=3) or female (n=3) participants – usually when conducted

in the Middle East. Their L1 background was also most commonly (n=40) that of the majority language of the country, but in five cases (S8, S11, S12, S14 and S19) they spoke this language, and it was utilised during 'L1'-mediated teaching, but some of them used others at home.

4.2.6: Outcome Measures and Study Findings

As stipulated by the eligibility criteria for this review, all 45 of the included studies employed quantitative measures of L2 vocabulary acquisition – generally in the form of either receptive definition identification tests, or productive L2 word provision tests. Altogether, more studies (n=20) found in favour of L1-mediated vocabulary teaching than L2-mediated (n=9), but eight observed no difference between the two, one reported unclear results on account of assessing multiple interventions simultaneously, and seven produced contradictory findings at post-test and delayed post-test. For a review of this and other key information, see Appendix 10.

4.3: Risk of Bias Results

Moving now to the risk of bias assessments that were conducted on all of the chosen studies, Table 11 presents both the individual ratings that were assigned to sections A-F of the EPHPP tool, and the global rating of each study as a whole. A score of 1 signifies that it was judged to be 'strong' in a particular area, 2 'moderate' and 3 'weak', and overall an equal amount (n=19) were labelled weak as moderate, and only eight as strong, as will be discussed below. Notice that the table reports on 46 assessments, as unlike the four experiments in S42 – which were analysed separately, but conducted together – the two in S36 were conducted separately.

Table 11: The Results of the Risk of Bias Assessment

	Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals and Drop-outs	Global Rating
Al Nassir (2012)	2	1	3	2	3	2	Weak
Al Ramadhan (2020)	2	2	3	2	3	3	Weak
Alcaraz-Mármol (2021)	2	2	1	2	2	1	Strong
Alsuhaymi (2019)	2	1	1	2	1	1	Strong
Amirian et al. (2016)	2	1	3	2	3	1	Weak
Arpacı (2016)	2	1	3	2	3	1	Weak
Arvizu (2020)	2	1	1	2	2	1	Strong
Baltova (1999)	1	1	2	2	2	1	Strong
Baranowska (2020)	2	2	1	2	3	1	Moderate
Bueno-Alastuey and García (2019)	2	2	1	2	3	3	Weak
Busse et al. (2020)	2	1	1	2	3	3	Weak
Busse et al. (2021)	2	1	1	2	3	3	Weak
Camó and Ballester (2015)	2	2	1	2	3	1	Moderate
Chalmers (2019)	3	1	1	2	3	1	Weak
Choi et al. (2016)	2	2	1	2	3	3	Weak
Del Pilar Agustín Llach (2009)	2	1	1	2	3	1	Moderate
Emirmustafaoglu and Gökmen (2015)	2	2	1	2	3	3	Weak
Ghabanchi and Ayoubi (2012)	2	1	1	2	3	3	Weak

Gyllstad et al. (2023)	3	2	1	2	2	1	Moderate
He (2019)	2	1	3	2	3	1	Weak
Helle (2014)	2	2	3	2	3	1	Weak
Hennebry et al. (2017)	1	1	2	2	3	1	Moderate
Hsu et al. (2013)	2	2	1	2	3	1	Moderate
Hu et al. (2014)	2	2	1	2	3	2	Moderate
Ina (2014)	2	1	1	2	3	1	Moderate
Kongtawee and Sappapan (2018)	1	2	3	2	3	3	Weak
Koolstra and Beentjes (1999)	2	1	1	2	2	1	Strong
Laufer and Girsai (2008a)	2	2	1	2	3	1	Moderate
Laufer and Girsai (2008b)	2	1	1	2	3	1	Moderate
Laufer and Shmueli (1997)	2	2	3	2	3	3	Weak
Lee and Macaro (2013)	1	2	1	2	3	1	Moderate
López (2016)	2	2	2	2	3	1	Moderate
Lwo and Lin (2012)	1	2	1	2	3	1	Moderate
Miyasako (2002)	2	1	3	2	2	1	Moderate
Muñoz et al. (2023)	2	1	1	2	3	2	Weak
Peters et al. (2016) (Experiment 1)	1	2	1	2	2	1	Strong
Peters et al. (2016) (Experiment 2)	2	2	1	2	2	2	Strong
Peters (2019)	2	1	1	2	2	1	Strong
Pujadas and Muñoz (2019)	2	1	1	2	3	2	Moderate
Ramachandran and Rahim (2004)	2	2	3	2	3	3	Weak
Sawada (2010)	2	2	1	2	2	3	Moderate

Sieh (2008)	2	2	3	2	2	1	Moderate
Tonzar et al. (2009) (Experiments 1-4)	2	2	3	2	3	3	Weak
Wu and Yang (2022)	2	1	3	2	3	1	Weak
Zhang and Graham (2020)	2	1	1	2	2	3	Moderate
Zhang (2022)	2	1	3	2	2	1	Moderate
Strong	6	23	29	0	1	28	8
Moderate	38	23	3	46	13	5	19
Weak	2	0	14	0	32	13	19

Note: A study was considered to have a 'strong' global rating if it had no weak classifications, 'moderate' if it had one weak classification, and 'weak' if it had two or more

4.3.1: Selection Bias

Beginning with the degree of selection bias in the included studies, most of them (n=38) were assessed as exhibiting moderate bias on the grounds that although they recruited participants who were 'very likely' to be representative of their target population, the percentage of them agreeing to take part went unstated. In six studies, the number consenting was stated – and as it was determined to be above 80%, they received a 'strong' rating, which denotes minimal bias – but unfortunately the remaining two studies (S14 and S19) were deemed to have a high risk of bias. This is because less than 60% of their chosen participants were ultimately involved (22.6% in S14 and 58.3% in S19), but it is important to note that while the EPHPP guidelines (see Appendix 11) advise a 'moderate' rating when this information is not provided, it is quite possible that some of the 'moderate' studies would have been deemed 'weak' if it were.

4.3.2: Study Design

Given that the random assignment of participants to experimental or control conditions is the most reliable means by which to ensure their equivalency prior to experimentation (Campbell & Stanley, 1963), the 23 RCTs in this review were all judged as possessing a 'strong' design. In spite of this, however, literally only one (S14) outlined the method of randomisation that was actually used – in this case, a random sequence generator on the website random.org – which

speaks to the ubiquity of underreporting in research. Regarding the 22 studies – containing 23 experiments – that were classified as NRCs, the EPHPP guidelines advocate that they also be rated as 'strong', but technically this advice pertains to the controlled clinical trials that were originally specified by the tool, and this review only comprises two categories of study, so the decision was made to reflect their weaker design by rating them as 'moderate' instead.

4.3.3: Confounders

While it is impossible in education research to control for all potential confounders, including those both unknown and unknowable, through any technique other than randomisation, this section of the EPHPP tool was reconceived to appraise the control only of known differences between the subject groups. As such, the researchers ought always to have been able to state whether their age, gender and socioeconomic status were equally distributed, but since these data were frequently unreported – which would have seen every study rated as 'weak' despite its control of other highly significant variables, such as L2 proficiency or pre-test scores – the approach taken was that studies were viewed as 'strong' (n=29) when the variables described were shown not to differ, or did differ, but were controlled in analysis. They were then viewed as 'moderate' (n=3) when merely 60-79% of confounders were controlled analytically, and as 'weak' (n=14) when less than 60% were controlled, or pre-tests were not administered.

4.3.4: Blinding

Insofar as the blinding of assessors to the intervention status of participants is concerned, this crucial piece of information went unshared in almost every study (n=44), with the only outlier being S14. In this research, both the individual who allocated subjects to conditions, and the assessor of their performance on pre-tests and post-tests, were unaware of their identity, but unfortunately, whether the participants knew the research questions was not confirmed, and so the study received a rating of 'moderate'. Of the other 44 studies, 26 of them also failed to verify participant awareness of the research questions, which should perhaps have seen them marked as 'weak', but the EPHPP guidelines generously allow for a 'moderate' assessment in this case as well. Finally, the 19 remaining studies did affirm that the research questions were concealed, but were still rated as 'moderate', so overall this area saw some risk of bias.

4.3.5: Data Collection Methods

Turning to the reliability and validity of the instruments employed to calculate L2 vocabulary acquisition in the 45 studies, those in as many as 32 of them were found to have a substantial risk of bias. The reason for this is that the researchers simply did not comment on their quality one way or the other, leading to 'weak' ratings, and on the occasions that they showed them to be reliable without establishing their validity (n=11), or valid without being reliable (n=1), they attained 'moderate' ratings. As evidence of validity, both correlations between test items and total test scores, and assessments of face, content or concurrent validity, were accepted, and as evidence of reliability, both assessments of inter-rater and internal consistency were accepted. Remarkably, only one study (S4) achieved a low risk of bias by demonstrating both of these, and this highlights the issue of untested instruments being used in research.

4.3.6: Withdrawals and Drop-Outs

Lastly, in terms of the amount of attrition in the included studies, the majority of them (n=28) were classified as exhibiting minimal bias because they recorded it clearly, and the percentage of participants completing the research ranged from 80-100%. In the case of four studies, the final percentage was from 60-79%, so they were given a rating of 'moderate', and then of the additional 13 studies – which were all considered to have a high risk of bias – twelve neglected to describe the presence or absence of attrition, and one (S12) clarified that three classes out of six were unable to participate, due to the coronavirus, but did not confirm their size.

4.4: Limitations of the Quality Assessment

All in all, the risk of bias assessment has revealed that the body of evidence applicable to this review is weak to moderate with respect to credibility, but problematically these global ratings do not convey the precise strengths and weaknesses of each individual study. In addition, the six sections of the EPHPP tool are not weighted, even though some – such as study design and blinding – are arguably more fundamental to the quality of research than others. If the former alone is considered, for example, then 23 of the included studies qualify as RCTs – albeit with cluster randomisation 69.6% of the time. If the latter is considered, then admittedly this factor is chronically underreported, but this relates to a third limitation of the assessment – namely that, although weak ratings for research that is poorly documented are wholly deserved, they do not necessarily reflect the reality of its organisation, implementation and analysis.

Chapter Five: Discussion

Now that a summary of the general findings of this review has been provided, this chapter will present a more detailed narrative synthesis of the interventions and outcomes of the selected studies, with the discussion structured according to the six categories of treatment identified in section 4.2.3. Altogether, of the 17 studies that judged in favour of L1-mediated vocabulary teaching, only one was granted a 'strong' global rating – while of the ten that decided against it, five achieved this rating – but this discovery acquires alternative significance as soon as the type of intervention is taken into account, not to mention the contradictory findings of eleven of the 45 studies. Notably, these studies were contradictory in the sense that the effects they ascertained, either in favour of L1 use (n=8) or against it (n=3), were unclear or inconsistent, and to view both this and the intervention type of the research visually, see Appendix 12.

5.1: Using L1 Translations or Pictures

Table 12: Summary of the Studies in Section 5.1

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Al Nassir (2012)	RCT	Secondary	Weak	Did not favour L1 use
Alsuhaymi (2019)	RCT	Primary	Strong	Did not favour L1 use
Bueno-Alastuey and García (2019)	NRC	Secondary	Weak	Did not favour L1 use
Emirmustafaoglu and Gökmen (2015)	NRC	Primary	Weak	Did not favour L1 use on the post-test, but no difference detected on the delayed post-test
Tonzar et al. (2009) (Experiments 1-4)	NRC	Primary	Weak	Did not favour L1 use

The first form of vocabulary teaching intervention to be evaluated in this systematic review is that of using L1 translations in comparison to pictorial representations to convey the meaning of target L2 words. While the latter approach may seem to be language neutral, it constitutes a L2-mediated method due to being employed alongside L2 vocabulary items – in either their spoken or written form – and Table 12 reveals that all of the research on this topic concluded in favour of picture-mediated teaching, despite its collectively high risk of bias. In a RCT at the secondary level, for example, Al Nassir (2012) taught 45 L2 words to 36 EFL learners, and those who were shown them beside pictures for thirty seconds acquired them to a greater degree

(*p*=0.043) than those who were shown them beside L1 translations. Admittedly, the study did not include a pre-test, and so the groups may have been variously familiar with the words, but a replication of it with 36 primary students (Alsuhaymi, 2019) yielded the same results.

This was with both a pre-test and a more protracted delayed post-test – at two weeks instead of four days – added to the procedure, and although both of these studies were conducted in Saudi Arabia, similar outcomes were obtained in other contexts as well. In their work with 52 secondary school pupils in Spain, for instance, Bueno-Alastuey and García (2019) determined that not only was picture-mediated vocabulary learning superior to the use of L1 translations, but it also outperformed video-mediated learning. Precisely what these three methodologies entailed they did not clearly explain – and statistical differences between the conditions were not calculated – but the superiority of pictures held true over time, and on both receptive and productive vocabulary tests. Promisingly, this echoed the findings of Tonzar et al. (2009), who taught 80 L2 English and German words to 229 Italian primary students, and observed images to aid their L2 vocabulary acquisition more than L1 translations even after one month.

In addition, this research also expanded our understanding of the potential limitations of the picture method, as it was found to be mediated by both the cognate status of the words, and the age of the participants. In other words, it was less successful than the translation method for the learning of cognate L2 words by 106 Grade 8 students, if not by 123 Grade 4 ones. With the exception of this cognate-specific discovery, however, the generally consistent findings of the four studies certainly suggest that pictures are preferable to L1 translation as a vocabulary teaching method most of the time, and the only caveat to this follows from Emirmustafaoglu and Gökmen (2015). This is because, while their implementation of the two approaches in the Turkish primary school context saw similar results when post-testing was compatible with the learning method, all of the subjects scored equally well when asked to label pictures. Perhaps, therefore, the apparent efficacy of picture-based learning is a product of biased testing.

5.2: Watching Television with L1 or L2 Audio and Subtitles

Table 13: Summary of the Studies in Section 5.2

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Baltova (1999)	RCT	Secondary	Strong	Did not favour L1 use
Baranowska (2020)	NRC	Secondary	Moderate	Did not favour L1 use
Camó and Ballester (2015)	NRC	Primary	Moderate	No difference detected on the post-test, but favoured L1 use on both delayed post-tests
Chalmers (2019)	RCT	Primary	Weak	No difference detected
He (2019)	RCT	Secondary	Weak	Favoured L1 use
Helle (2014)	NRC	Secondary	Weak	No difference detected
Hsu et al. (2013)	NRC	Primary	Moderate	Favoured L1 use
Ina (2014)	RCT	Primary	Moderate	Favoured L1 use
Koolstra and Beentjes (1999)	RCT	Primary	Strong	Favoured L1 use
Lwo and Lin (2012)	NRC	Secondary	Moderate	No difference detected
Muñoz et al. (2023)	RCT	Secondary	Weak	No difference detected
Peters et al. (2016) (Experiments 1-2)	NRC	Secondary	Strong	Did not favour L1 use
Peters (2019)	RCT	Secondary	Strong	Did not favour L1 use
Pujadas and Muñoz (2019)	RCT	Secondary	Moderate	No difference detected
Wu and Yang (2022)	RCT	Secondary	Weak	Favoured L1 use on two post-tests, but no difference detected on the third post-test

Moving to the most common kind of vocabulary teaching intervention unveiled in this review, Table 13 illustrates that although most of the studies exhibited a low (n=4) to moderate (n=6) risk of bias, the inconsistency of their conclusions was nonetheless considerable. In sum, after all, five of them favoured L1 use while four did not, and another six either found no significant differences between L1-mediated and L2-mediated learning, or detected it on some post-tests only. Furthermore, just two of the studies (Baltova, 1999; Camó & Ballester, 2015) performed pre-tests, post-tests and delayed post-tests, with seven assuming a pre-test/post-test design, and six administering post-tests only, so the strength of the research in this respect is limited as well. To provide an overview of the specific treatments investigated in these 15 studies, the following two sections will describe first those (n=8) that focused on incidental L2 vocabulary acquisition, and then those (n=7) that focused on intentional L2 vocabulary acquisition.

5.2.1: Studies on Incidental Vocabulary Learning

Of the eight studies on this topic, which Hulstijn (2013) defines as the unintended attainment of L2 words through exposure to linguistic stimuli, all of them explored it by exposing learners to various combinations of L1 or L2 audio and subtitles. With regard to the two at the primary level, for instance, participants watched either a 12-minute television programme about bears (Koolstra & Beentjes, 1999), or a 15-minute cartoon (Ina, 2014) with L1 audio, L2 audio or L2 audio and L1 subtitles. The procedure of the latter was based on the procedure of the former, with the students receiving the input in small groups, and in both cases the subtitled viewing proved the most effective for the acquisition of 28 or 35 target L2 words, respectively. On the subject of limitations, each study assessed only receptive vocabulary knowledge, and utilised the L1 audio condition as a source of comparison, in lieu of more accurate pre-tests, but they were RCTs, with large sample sizes, robust risk of bias ratings, and equivalent outcomes.

At the secondary level, a study with the same design that exposed 123 Norwegian L1 speakers and 89 German L1 speakers to a 20-minute clip from an English language film failed to detect any meaningful differences between the conditions, however. Helle (2014) basically observed that the Norwegian and Austrian participants performed much the same in all conditions, but importantly, she did not calculate inferential statistics, and the study was judged as 'weak' on account of not controlling for the confounders of L2 and even L1 proficiency. A study with two experiments (Peters et al., 2016) with 28 EFL learners from an academic secondary school and 18 EFL learners from a vocational secondary school then altered this design by comparing the effects of L2 audio and L1 subtitles with L2 audio and L2 captions, and on both occasions the latter enabled greater learning of L2 word form, and equal learning of L2 word meaning.

In other words, it showed L2 captions to be more beneficial for incidental vocabulary learning than L1 subtitles – with both intermediate and elementary learners – and this was significant as the rating of the study was strong. That said, neither its design nor its scale were ideal, and so to improve upon this, Peters (2019) conducted a third experiment with 118 Dutch-speaking secondary students, and this time it was a RCT with an additional control group that exposed them to a 12-minute documentary with L2 audio only. In this study, L2 audio with L2 captions were once again shown to facilitate superior lexical gains in terms of (1) form recognition, and (2) meaning recall, and this outcome reproduced that of earlier research with the same three

conditions (Baranowska, 2020). As these studies achieved a low to moderate risk of bias, their findings are compelling, but it is worth noting that they assessed only receptive knowledge of the target L2 words, and that participants had their focus drawn to them by pre-tests.

Turning, then, to a similar study without a pre-test, Lwo and Lin (2012) asked 32 Mandarin L1 speakers to undertake two L2 lessons using a multimedia reading program that enabled them to hear and read text simultaneously. On top of the three conditions of the prior studies, they added a fourth wherein subjects experienced L2 audio with both L1 and L2 captions, and their receptive test of incidental L2 vocabulary acquisition following the intervention uncovered no differences between the groups. While it did not reach statistical significance, there was some benefit for low level learners of L2 audio with both L2 captions, and L1+L2 captions, however, and a 'strong' study by Baltova (1999) that uniquely compared L2 audio and L2 audio with L2 captions to L1 audio with L2 captions also concluded that the second allowed for significantly better vocabulary learning. This included retention, since it was the sole study to administer a delayed post-test, and so overall the research in this section did not benefit from L1 use.

5.2.2: Studies on Intentional Vocabulary Learning

Unlike the preceding section, where seven of the studies were deemed 'moderate' or 'strong' and only one 'weak', those in this section have a much higher risk of bias, with three regarded as 'moderate' and four as 'weak'. They are also not as cohesive, with intentional L2 vocabulary acquisition explored in several ways, and to start with research at the primary level, Camó and Ballester (2015) showed 34 Grade 5 EFL learners a 10-minute cartoon with pictorial flashcards of the target L2 words, and either L1 oral translations of them, or no translations. Intriguingly, they noticed no differences between conditions on a post-test of receptive lexical knowledge, but significant differences in favour of L1 use on four-day and one-month delayed post-tests, and Chalmers (2014) echoed this when he also detected no differences between EAL learners who watched video-based explanations of target L2 words in their L1 or L2. Unfortunately, he did not conduct delayed post-tests, but he did assess receptive and productive learning.

Potentially, then, L1-mediated and L2-mediated vocabulary teaching may be equally effective insofar as short-term intentional learning is concerned – if with varied effects over time – but research into another type of intervention intimates otherwise. Hsu et al. (2013), for example,

compared viewing four 20-second educational videos with L2 audio, L2 audio and L2 captions with L1 translations of the target words, and L2 audio with L1+L2 captions of the target words only, and found the latter two to be equally beneficial for receptive vocabulary learning. This was with primary students, but in a small-scale pilot study with just six secondary learners, He (2019) discovered L1 target word captioning to be superior to L2 target word captioning or no captioning when seeing L2 videos for productive vocabulary learning too. Lastly, Wu and Yang (2022) played a TED Talk for 133 high school students in four conditions, and also established that L2 audio with L2 captions and L1 target word translations best assisted acquisition.

None of these studies directly compared L2 audio and L2 captions, which were observed to be advantageous for incidental L2 vocabulary acquisition in the previous section, to L2 audio and L2 captions with L1 target word translations – to see whether incidental learning may outclass intentional learning – but one study did compare them to both L2 audio with L1 subtitles, and these conditions with pre-viewing and post-viewing vocabulary tasks. Ultimately, Pujadas and Muñoz (2019) concluded that their 74 Grade 8 participants recalled the form and meaning of 120 target L2 words significantly better with the support of such tasks – and regardless of the language of transcription – and in a follow-up study (Muñoz et al., 2023), L2 audio with either L1 or L2 captions and post-viewing vocabulary tasks were again compared, and again found to be similarly effective. In other words, intentional L2 vocabulary acquisition seems to be aided by captioning type only if word form, meaning and use will not be additionally practised.

5.3: Listening with L1 or L2 Glossing, Storytelling or Contrastive Analysis

Table 14: Summary of the Studies in Section 5.3

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Amirian et al. (2016)	RCT	Secondary	Weak	No difference detected
Arvizu (2020)	RCT	Primary	Strong	May have favoured L1 use, but unclear because multiple interventions were tested concurrently
Hennebry et al. (2017)	RCT	Secondary	Moderate	Favoured L1 use
López (2016)	NRC	Secondary	Moderate	Did not favour L1 use
Sawada (2009)	NRC	Secondary	Moderate	Favoured L1 use on one of the post-tests

				and delayed post-tests, but no difference detected on the other
Sieh (2008)	NRC	Primary	Moderate	No difference detected on Story A, but favoured L1 use on Story B
Zhang and Graham (2020)	RCT	Secondary	Moderate	Favoured L1 use
Zhang (2022)	RCT	Secondary	Moderate	Favoured L1 use

The third kind of vocabulary teaching intervention identified by this review is that of listening to L2 texts, and clarifying the meaning of target words in various L1-mediated or L2-mediated ways. There are three categories of pedagogical approach in all, and the studies in this section are generally sound, with Table 14 demonstrating that seven of them attained a 'strong' (n=1) or 'moderate' (n=6) quality rating, and that only one was assessed as 'weak'. Their inferences are relatively consistent too, seeing as one ruled against L1 use, and one judged it identical to L2 use, but six decided in favour of it – or at least produced some evidence to that effect. Half of the studies adopted a robust pre-test/post-test/delayed post-test design as well, with none of them administering post-tests only, which strengthens the credibility of their findings.

5.3.1: Studies on Oral or Written Glossing

Beginning with the only study on this topic with primary learners, Sieh (2008) asked 63 Grade 4 EFL students to listen to two L2 stories with pictorial flashcards of their target L2 words, and either L1 voiced translations of them, or no translations. For Story B, there were no significant differences between the conditions at pre-test, post-test or delayed post-test, but for Story A, those disadvantaging the L1-mediated group in the pre-test disappeared in subsequent tests, which implies that they both acquired and retained more L2 words. That said, Sieh (2008) only assessed passive recall of their meaning – by asking learners to match them to pictures – so in research that builds on this at the secondary level, Sawada (2009) evaluated both passive and active recall of 45 target L2 words. This was after 116 Japanese participants heard L1, L2 or no explanations of them while listening to nine L2 texts, and overall the first two conditions were equally effective for passive recall, but the former much more effective for active recall.

This was again on post-tests and delayed post-tests, so there is some justification for the claim that L1 glossing facilitates receptive L2 vocabulary learning more than L2 glossing, but a study

with 74 Iranian high school students casts doubt on this. It does so by revealing that when 96 target L2 words in eight L2 stories were orally defined for them in either their L1, L2 or L1 and L2, each method had a similar effect on short-term and long-term acquisition. Admittedly, this may be because scores on tests of passive recall, active recall and active use were combined, or because Amirian et al. (2016) did not conduct pre-tests, but another study by López (2016) that investigated the use of L1 translations, L2 definitions or no definitions of target L2 words during oral storytelling also found the L1 to be non-beneficial. That is, its use produced lower scores on checks of both receptive and productive vocabulary knowledge than the L2 did, and although this effect lessened as L2 proficiency increased, it endured across ability levels.

In sum, what this means is that the findings of research on this vocabulary teaching approach are highly conflicting – perhaps on account of the studies attaining an insufficiently low risk of bias – and the final experiment to be discussed in this section only exacerbates this point. This is because Hennebry et al. (2017) concluded in favour of L1 use after exposing 262 secondary students of French to the same interventions as López (2016), but with regard to 60 target L2 words over eight L2 texts, instead of five target L2 words in just one text. They also showed it to help low proficiency and high proficiency learners equally, and while the increased scale of the research lends its greater authority – as does the fact that it employs a crossover AB/BA design – regrettably only receptive L2 vocabulary knowledge was assessed, and lexical glosses were given in writing, rather than orally. Since either of these points could be responsible for its alternative outcomes, it seems clear that more directly analogous research is needed.

5.3.2: Studies on Oral Storytelling

Proceeding to the only study in this section that utilised this kind of intervention, Arvizu (2020) undertook a RCT with 167 Mexican primary students from Grades 3, 4, 5 and 6 that achieved a ‘strong’ quality evaluation rating, and exposed them to 24 target L2 words in twelve spoken L2 stories. Pupils were either (1) pre-taught the target words by means of pictorial flashcards, a contextualising L1 story, and a class discussion of their meaning – as well as (2) requested to produce them in a paired retelling of each story – or they were not, and receptive vocabulary acquisition was measured. Interestingly, while Arvizu (2020) detected no difference between the conditions on the post-test, a small one was found in favour of the experimental group on a four-month delayed test, which implies that the treatment had a long-term impact. With an

effect size as minor as *r*=0.19, however, and so many vocabulary teaching methods evaluated at once, it is impossible know what caused this impact, or whether the L1 was useful at all.

5.3.3: Studies on Contrastive Analysis

Finally, in contravention of the inconclusive results of the foregoing sections, two studies that enacted a contrastive focus on L1 and L2 word form during L2 vocabulary instruction in one of their conditions mutually determined in support of it. In the first of these, Zhang and Graham (2020) compared providing L1 definitions and example L2 sentences for 60 target L2 words in six L2 oral texts with providing L2 definitions and example L2 sentences, L1 and L2 definitions with contrastive focus-on-form, or no definitions. They enlisted 137 Chinese EFL learners, and in the second study – which expanded this research – Zhang (2022) then compared providing 98 of the same secondary students with L1 or L2 definitions and example L2 sentences for 20 target L2 words, or L1 definitions and contrastive focus-on-form. In both experiments, greater gains in active recall of L2 word meanings occurred when L1/L2 differences were highlighted, and so it may be that L2 vocabulary acquisition is best mediated through both languages.

5.4: Reading with L1 or L2 Glossing or Writing Activities

Table 15: Summary of the Studies in Section 5.4

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Al Ramadhan (2020)	NRC	Secondary	Weak	Favoured L1 use
Alcaraz-Mármol (2021)	NRC	Primary	Strong	Did not favour L1 use
Arpacı (2016)	RCT	Secondary	Weak	Favoured L1 use
Choi et al. (2016)	NRC	Secondary	Weak	No difference detected on the post-test, but partly favoured L1 use on the delayed post-test
Del Pilar Agustín Llach (2009)	RCT	Secondary	Moderate	Favoured L1 use on both of the post-tests, but no difference on the delayed post-tests
Ghabanchi and Ayoubi (2012)	RCT	Secondary	Weak	No difference detected
Gyllstad et al. (2023)	NRC	Secondary	Moderate	Favoured L1 use on the post-test, but no difference detected on the delayed post-test

Hu et al. (2014)	NRC	Secondary	Moderate	No difference detected on the post-tests, but did not favour L1 use for high-level learners on one delayed post-test
Kongtawee and Sappapan (2018)	NRC	Secondary	Weak	Favoured L1 use
Laufer and Shmueli (1997)	NRC	Secondary	Weak	Favoured L1 use
Lee and Macaro (2013)	NRC	Primary	Moderate	Favoured L1 use
Miyasako (2002)	RCT	Secondary	Moderate	Did not favour L1 use on the post-test, but no difference detected on the delayed post-test
Ramachandran and Rahim (2004)	NRC	Secondary	Weak	Favoured L1 use

In this section, the next kind of vocabulary teaching approach to be evaluated shifts the focus from hearing L2 texts to reading them, and again moderating the comprehension of target L2 words through either the L1 or the L2. Their outcomes are some of the most consistent in the review, with nine out of thirteen studies electing either in favour of L1 use (n=7) or potentially in favour of it (n=2), and according to Table 15, only two actively settled against it. Even then, their findings on one occasion were transient, but the cogency of this coherence is weakened by the high (n=7) or moderate (n=5) risk of bias that the research exhibited, and the drawback that only one paper received a 'strong' quality rating. This is despite the fact that most of the studies (n=12) took the time to administer delayed post-tests, and so were structurally sound, and the following sections will review first those that investigated various forms of written or oral glossing, and then those that amalgamated written glossing with writing activities.

5.4.1: Studies on Written or Oral Glossing

To start with the experiments – of which there are seven – that studied the effects of viewing written definitions of target L2 words while reading on L2 vocabulary growth, the earliest one (Hu et al., 2014) concentrated on optional viewing. More specifically, it directed 78 high-level and low-level Taiwanese EFL learners to read two L2 articles on bespoke online software, and willingly expose themselves to either L1 translations of the 12 target L2 words in Article A, or L2 definitions of the 12 target L2 words in Article C. This was enabled by allowing them to click on hypertext glosses whenever they chose, and although no differences were found between

the groups on measures of receptive and productive vocabulary knowledge at post-test – and only a slight indication that L2 glosses may be more effective for high-level subjects regarding the retention of receptive word meanings at delayed post-test – this may well be because the frequency with which they accessed the glosses was neither mandated nor monitored.

Indeed, in later research (Arpacı, 2016; Choi et al., 2016; Kongtawee & Sappapan, 2018) where participants were required to access printed or computerised L1 or L2 glosses – in addition to no glosses, in the first two cases – outcomes were clearly in support of L1 use. This may have been with only borderline significance in the first study, for only the retention of oft-occurring words in the second study, and for only receptive L2 vocabulary knowledge in every study, but they were still congruous determinations. When more complex forms of written glossing were explored, such as seeing L1 translations or L2 definitions of target L2 words alongside sample L2 sentences, paragraphs or elaborated texts (Laufer & Shmueli, 1997), or L1 translations both in comparison to and in conjunction with L2 word repetition (Al Ramadhan, 2020), L1 glossing was again shown to facilitate superior L2 vocabulary acquisition, and to do so both receptively in terms of meaning recognition in the first study, and productively in the second study.

In fact, the only research to judge against L1-mediated glossing techniques on a post-test was Miyasako (2002), who compared L1 multiple-choice or single glossing with L2 multiple-choice or single glossing, no glossing or no reading conditions. The 187 Japanese secondary students in this study were subjected to these conditions – except in the latter case – while answering comprehension questions on a L2 reading passage, and tests of their receptive understanding of 15 target L2 words showed that acquisition was best assisted by multiple-choice glossing in their L2. That said, however, this method was only significantly better in comparison to the no glossing groups, appeared to benefit only high-level learners, and did not maintain its efficacy on an 18-day delayed post-test, so the finding is somewhat tenuous. Furthermore, in the only study with primary learners in this section, Lee and Macaro (2013) discovered oral L1 glossing while reading to better aid receptive lexical learning, so L1 glosses seem largely effective.

5.4.2: Studies on Written and Oral Glossing

Continuing with research that exposed EFL learners to both written and oral glosses during L2 reading, Ramachandran and Rahim (2004) did this by (1) writing on the board, and (2) verbally

explaining the L1 or L2 definitions of 20 target L2 words in four L2 texts. They did this with 60 Malaysian participants, and ascertained that those in the L1 glossing condition outperformed their peers to a highly significant degree both immediately, and after a one-month delay. This was with regard to active recall of their meanings, and in research with 74 secondary students in Sweden by Gyllstad et al. (2023), this same aspect of receptive knowledge was again found to be assisted by written and oral L1 glossing. In this case, such glossing involved utilisation of the majority language of Swedish in one condition, and all other L1s known to the participants in a second, and the researchers established that both generated higher scores – with a small effect size – relative to L2-mediated teaching, but that this effect waned 8-10 weeks later.

Whether L1 glosses expedite long-term L2 vocabulary acquisition when they are both written and spoken thus remains unclear, with too few studies conducted on the topic, but it is worth noting that an additional experiment was undertaken at the primary level by Alcaraz-Mármol (2021). Notably, this was the only study to receive a 'strong' quality rating in this section, and rather than analysing the use of written and spoken L1 or L2 glosses in combination, it instead contrasted the two, with 20 Spanish-speaking students in one condition reading three L2 texts with L1 translations of 10 target L2 words provided in the margins, and 20 students in another condition reading these texts out loud, and undertaking 'dynamic' glossing. What this means is that when they came to target words, the researcher offered them up to five L2 prompts to help convey their meaning, and tests of both receptive and productive L2 knowledge revealed that this inductive method was the most effective of the two, in the short and long-term.

5.4.3: Studies on Written Glossing and Writing Activities

Moving in this last section to research that united L1 or L2 written glossing while reading with L2 writing activities, there were more uniform results, with Del Pilar Agustín Llach (2009) and Ghabanchi and Ayoubi (2012) ultimately finding no difference between their conditions. In the first study, for example, 27 German learners of L2 Spanish either viewed 10 target L2 words in a bilingual list, and wrote L2 sentences containing them, or they viewed them in a L2 text with L1 translations in one condition, and access to a dictionary in another. Post-tests of receptive and productive L2 vocabulary acquisition detected a significant effect in favour of the writing task, but this disappeared after three weeks, and in the second study, 120 Iranian participants were instructed to read two L2 passages with L1 glosses of the 30 target L2 words, L2 glosses,

a dictionary, or L2 glosses while writing a L2 summary of their content. Once again, the writing task yielded higher scores at post-test and delayed post-test, but this was not significant.

5.5: Reading with Translation Activities, Content Discussion or Vocabulary Tasks

Table 16: Summary of the Studies in Section 5.5

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Laufer and Girsai (2008a)	NRC	Secondary	Moderate	Favoured L1 use
Laufer and Girsai (2008b)	RCT	Secondary	Moderate	Favoured L1 use

The penultimate type of vocabulary teaching intervention to be included in this review is that of reading L2 texts while engaging in translation activities, content discussion, or post-reading vocabulary tasks, with Laufer and Girsai (2008a) comparing the first of these. In their research, 48 Israeli secondary students were tasked with completing true-or-false questions about a L2 text, and then either discussing its contents, or translating it both into and from their L1. When assessed on their acquisition of 17 target L2 words, those in the latter group were determined to have learned significantly more both receptively and productively, and in Laufer and Girsai (2008b), 26 Israeli participants who responded to questions about a L2 text before translating it twice were again observed to have learned significantly more of the 10 target L2 words. This was in comparison to both content debate and vocabulary tasks, and the result held true on a delayed post-test, so overall L1 use in the form of two-way translation appears beneficial.

5.6: Using Translation-Based or Textbook-Based Learning Activities

Table 17: Summary of the Studies in Section 5.6

Study Reference	Study Design	Study Context	Risk of Bias Assessment	Study Finding
Busse et al. (2020)	RCT	Primary	Weak	Favoured L1 use
Busse et al. (2021)	RCT	Primary	Weak	Favoured L1 use

Given the finding of the previous section, the focus of the final one on whether L2 vocabulary acquisition is best facilitated by translation-based or textbook-based tasks is opportune, even if the studies were rated as ‘weak’. In each of them, multilingual German speakers were taught

the same target L2 words over five 45-minute lessons, and Busse et al. (2020) explored doing so with an emphasis on L1 equivalents in one condition, or textbook-based lexical exercises in the other. Busse et al. (2021) then elaborated on this by inserting a third condition that aimed to develop positive attitudes towards plurilingualism, and in both cases they gauged receptive and productive vocabulary acquisition and retention using the same tests. They also produced the same results, with both studies finding the generation of translations in a range of L1s to enable significantly better learning – including at a four-week and 2.5-week interval – and this was despite the fact that learners in these groups spent less time practicing the L2 words.

Chapter Six: Conclusion

6.1: Summary of the Findings

In sum, this review has located 45 studies, under six broad categories of intervention type, on the effects of using the L1 in comparison to the L2 to teach target language lexis in primary or secondary school settings. It did so by way of a meticulous searching, screening and selection process in order to gain evidence-based insight into the validity of claims that multilingualism should be utilised as a resource in L2 learning, and following the detailed examination of each included study in Chapter 5, an overview of its general findings is presented in Table 18.

Table 18: Summary of Findings by Intervention Type

	Intervention Type and Overall Finding
5.1	**Using L1 Translations or Pictures** Overall Finding: Appeared to favour the use of pictures, but with some indication that this may be the result of condition-congruent testing
5.2	**Watching Television with L1 or L2 Audio and Subtitles**
	Studies on Incidental Vocabulary Learning Overall Finding: Favoured the use of L2 audio with L2 captions over L2 audio with L1 subtitles, or L2 audio/L1 audio only
	Studies on Intentional Vocabulary Learning Overall Finding: Favoured the use of L2 audio with L1 target word translations – and possibly L2 captions, although this is less clear – or L2 audio with L1/L2 captions and additional vocabulary tasks
5.3	**Listening with L1 or L2 Glossing, Storytelling or Contrastive Analysis**
	Studies on Oral or Written Glossing Overall Finding: Tentatively favoured the use of L1 oral glossing, but with some indication that L2 oral glossing may also be effective

	Studies on Oral Storytelling Overall Finding: Slight indication that pre-teaching target L2 words by means of a contextualising L1 story has long-term benefits, but unclear as multiple interventions were implemented simultaneously
	Studies on Contrastive Analysis Overall Finding: Favoured the use of a contrastive focus on L1 and L2 equivalent word forms, along with either L1 or L2 oral glossing
5.4	**Reading with L1 or L2 Glossing or Writing Activities**
	Studies on Written or Oral Glossing Overall Finding: Favoured the use of written or oral L1 glossing, both alone, and with sample L2 sentences and target L2 word repetition
	Studies on Written and Oral Glossing Overall Finding: Tentatively favoured written and oral L1 glossing in combination, and inductive learning through oral L2 glossing
	Studies on Written Glossing and Writing Activities Overall Finding: No difference between written L1/L2 glossing and written L1/L2 glossing with target L2 word writing activities
5.5	**Reading with Translation Activities, Content Discussion or Vocabulary Tasks** Overall Finding: Favoured forwards and backwards translation of a L2 text over content discussion and post-reading vocabulary tasks
5.6	**Using Translation-Based or Textbook-Based Learning Activities** Overall Finding: Favoured learning target L2 words alongside L1 oral or written equivalents, rather than through textbook-based vocabulary exercises or the promotion of positive plurilingual attitudes

6.2: Limitations of the Review

While the findings summarised above can provide tentative guidance to L2 teachers, students and policy-makers, it would be inaccurate to regard them as recommendations per se. In part, this is because there are so few studies within each intervention type that outcomes have not been suitably replicated, and in part it is because of their predominantly poor methodological quality, with 38 of them rated as 'moderate' or 'low'. Not only this, but it is quite possible that the study identification process may have failed to detect all relevant research, since no hand screening of paper journals was conducted, and search terms were in English only. This means that any studies in languages other than English whose abstracts – or at least keywords – had not been translated would go unseen, and this likely applies especially to grey literature.

Publication bias, which is defined by Siddaway et al. (2019) as the tendency to publish research only when it contains a statistically significant result, is another pervasive issue that doubtless affected this review, and so too is the fact that it was directed primarily by one reviewer. This

is despite advice (Macaro, 2019) for all systematic reviews to be undertaken by a team where possible, and while attempts were made to minimise subjectivity in this research, only 10% of study abstracts and full-texts were double-screened, and the processes of data extraction and synthesis were both done by only one person. Finally, it must be noted that although unclear or incomplete reporting of research was rife, there was insufficient time to contact all authors for clarification of their work, and interpretation of it was necessarily constrained by this.

6.3: Recommendations for Future Research

With the exception, perhaps, of research on L2 vocabulary learning while watching television, which is relatively robust, this review has underscored the need for additional research to be conducted in all five of the other areas identified. Ideally, this would be on a larger scale and with the guaranteed inclusion of both pre-tests and control groups, but for now, the studies appraised in this review have provided tentative support for the advantages of L1 use.

Reference List

Andersen, R. (1976). A functional acquisition hierarchy study in Puerto Rico. In *10th Annual TESOL Convention, New York*.

Antón, M., & DiCamilla, F. (1998). Socio-cognitive functions of L1 collaborative interaction in the L2 classroom. *Canadian Modern Language Review*, *54*(3), 314-342. https://doi.org/10.3138/cmlr.54.3.314

Armijo-Olivo, S., Stiles, C., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: Methodological research. *Journal of Evaluation in Clinical Practice, 18*(1), 12-18. https://doi.org/10.1111/j.1365-2753.2010.01516.x

Atkinson, D. (1987). The mother tongue in the classroom: A neglected resource? *ELT Journal*, *41*(4), 241-247. https://doi.org/10.1093/elt/41.4.241

Atwill, K., Blanchard, J., Gorin, J. S., & Burstein, K. (2007). Receptive vocabulary and cross-language transfer of phonemic awareness in kindergarten children. *The Journal of Educational Research*, *100*(6), 336-346. https://doi.org/10.3200/joer.100.6.336-346

Auerbach, E. R. (1993). Reexamining English only in the ESL classroom. *TESOL Quarterly, 27*(1), 9-32. https://doi.org/10.2307/3586949

Baker, C. (2006). *Foundations of bilingual education and bilingualism.* Multilingual Matters.

Barnett, W. S., Yarosz, D. J., Thomas, J., Jung, K., & Blanco, D. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly*, *22*(3), 277-293. https://doi.org/10.1016/j.ecresq.2007.03.003

Bayley, S. N. (1998). The Direct Method and modern language teaching in England 1880-1918. *History of Education*, *27*(1), 39-57.

https://doi.org/10.1080/0046760980270104

Benson, C. (2002). Transfer/Cross-linguistic influence. *ELT Journal*, *56*(1), 68-70. https://doi.org/10.1093/elt/56.1.68

Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading*, *9*(1), 43-61. https://doi.org/10.1207/s1532799xssr0901_4

Bialystok, E., McBride-Chang, C., & Luk, G. (2005). Bilingualism, language proficiency, and learning to read in two writing systems. *Journal of Educational Psychology*, *97*(4), 580-590. https://doi.org/10.1037/0022-0663.97.4.580

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*(2), 384-414. https://doi.org/10.2307/414678

Brooks, F. B., & Donato, R. (1994). Vygotskyan approaches to understanding foreign language learner discourse during communicative tasks. *Hispania*, *77*(2), 262-274. https://doi.org/10.2307/344508

Brown, M., & Richardson, M. (2017). Understanding and synthesizing numerical data from intervention studies. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 131-154). SAGE Publications.

Brunton, G., Stansfield, C., Caird, J., & Thomas, J. (2017). Finding relevant studies. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 93-122). SAGE Publications.

Burstall, C., Jamieson, M., Cohen, S. & Hargreaves, M. (1974). *Primary French in the Balance*. National Foundation for Educational Research in England and Wales (NFER).

Burt, M., & Dulay, H. (1978). Some guidelines for the assessment of oral language proficiency and dominance. *TESOL Quarterly, 12*(2), 177-192. https://doi.org/10.2307/3585608

Butzkamm, W. (2003). We only learn language once. The role of the mother tongue in FL classrooms: Death of a dogma. *The Language Learning Journal*, *28*(1), 29-39.

https://doi.org/10.1080/09571730385200181

Butzkamm, W. (2007). Native language skills as a foundation for foreign language learning. In W. Kindermann (Ed.), *Transcending boundaries. Essays in honour of Gisela Hermann-Brennecke* (pp.71-85). Lit Verlag.

Campbell Collaboration. (n.d.). *Campbell Collaboration Online Library*. Campbell Collaboration. Retrieved January 28, 2023, from https://www.campbellcollaboration.org/component/jak2filter/?Itemid=1352&issearch=1&isc=1&category_id=101&xf_4[0]=2&xf_8[0]=3&ordering=publishUp

Campbell, D., & Stanley, J. (1963). *Experimental and quasi experimental designs for research*. Houghton Mifflin Co.

Carey, S. T., & Cummins, J. (1979). English and French achievement of grade 5 children from English and mixed French-English home backgrounds attending the Edmonton separate school system English-French immersion program. *Report submitted to the Edmonton separate school system.*

Celic, C., & Seltzer, K. (2012). Translanguaging: A CUNY-NYSIEB guide for educators. *CUNY-NYSIEB.*

Cenoz, J., & Gorter, D. (2021). *Pedagogical translanguaging*. Cambridge University Press. https://doi.org/10.1017/9781009029384

Chalmers, H. (2014). *Harnessing linguistic diversity in polylingual British curriculum schools. Do L1 mediated home learning tasks improve learning outcomes for bilingual children? A randomised trial.* [Master's dissertation, Oxford Brookes University]. British Council.

Chalmers, H. (2019). *Leveraging the L1: The role of EAL learners' first language in their acquisition of English vocabulary* [Unpublished doctoral thesis, Oxford Brookes University]. https://doi.org/10.13140/RG.2.2.20764.13445

Chalmers, H., & Murphy, V. (2022). Multilingual learners, linguistic pluralism and implications

for education and research. In E. Macaro & R. Woore (Eds.), *Debates in second language education* (pp. 66-88). Routledge. https://doi.org/10.4324/9781003008361-6

Chuang, H.-K., Joshi, R. M., & Dixon, L. Q. (2011). Cross-language transfer of reading ability. *Journal of Literacy Research*, *44*(1), 97-119. https://doi.org/10.1177/1086296x11431157

Chumak-Horbatsch, R. (2012). *Linguistically appropriate practice: A guide for working with young immigrant children*. University Of Toronto Press.

Codina Camó, A., & Pladevall Ballester, E. (2015). The effects of using L1 translation on young learners' foreign language vocabulary learning. *Elia*, *15*(1), 109-134. https://doi.org/10.12795/elia.2015.i15.06

Cohen, A. D. (1974). The culver city Spanish immersion program: The first two years. *The Modern Language Journal*, *58*(3), 95-103.

Comeau, L., Cormier, P., Grandmaison, É., & Lacroix, D. (1999). A longitudinal study of phonological processing skills in children learning to read in a second language. *Journal of Educational Psychology*, *91*(1), 29-43. https://doi.org/10.1037/0022-0663.91.1.29

Conteh, J., & Meier, G. (2014). *The multilingual turn in languages education*. Multilingual Matters.

Cook, V. (1992). Evidence for multicompetence. *Language Learning*, *42*(4), 557-591. https://doi.org/10.1111/j.1467-1770.1992.tb01044.x

Cook, V. (1995). Multi-competence and the learning of many languages. *Language, Culture and Curriculum*, *8*(2), 93-98. https://doi.org/10.1080/07908319509525193

Cook, V. (2001). Using the first language in the classroom. *Canadian Modern Language Review*, *57*(3), 402-423. https://doi.org/10.3138/cmlr.57.3.402

Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The Modern Language Journal*, *94*(1), 103-115. https://doi.org/10.1111/j.1540-4781.2009.00986.x

Cummins, J. (1976). *The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. Working papers on bilingualism, No. 9.* Bilingual Education Project, The Ontario Institute for Studies in Education.

Cummins, J. (1977). A comparison of reading achievement in Irish and English medium schools. In V. Greaney (Ed.), *Studies in Reading*. Educational Co. of Ireland.

Cummins, J. (1979a). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222-251. https://doi.org/10.2307/1169960

Cummins, J. (1979b). *Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. Working papers on bilingualism, No. 19.* Bilingual Education Project, The Ontario Institute for Studies in Education.

Cummins, J. (1980a). The construct of language proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown university round table on languages and linguistics 1980* (pp. 81-103). Georgetown University Press.

Cummins, J. (1980b). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, *14*(2), 175-187. https://doi.org/10.2307/3586312

Cummins, J. (1981a). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework* (pp. 3-49). Office Of Bilingual Bicultural Education. http://dx.doi.org/10.13140/2.1.1334.9449

Cummins, J. (1981b). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, *2*(2), 132-149.

Cummins, J. (1998). Immersion education for the millennium: What have we learned from 30 years of research on second language immersion? In M. R. Childs & R. M. Bostwick (Eds.), *Learning through two languages: Research and practice*. *Second Katoh Gakuen International Symposium on Immersion and Bilingual Education* (pp. 34-47). Katoh Gakuen, Japan.

Cummins, J. (1999). *BICS and CALP: Clarifying the distinction.* University of Toronto.

Cummins, J. (2000). Critiques of the conversational/academic language proficiency distinction. In J. Cummins (Ed.), *Language, power and pedagogy: Bilingual children in the crossfire* (pp. 86-111). Multilingual Matters. https://doi.org/10.21832/9781853596773-005

Cummins, J. (2001). Bilingual children's mother tongue: Why is it important for education? *Sprogforum*, *19*(1), 15-20.

Cummins, J. (2007). Rethinking monolingual instructional strategies in multilingual classrooms. *Canadian Journal of Applied Linguistics*, *10*(2), 221-240.

Cummins, J. (2008). Teaching for transfer: Challenging the two solitudes assumption in bilingual education. *Encyclopedia of Language and Education*, *5*(1), 1528-1538.

Cummins, J. (2016). Reflections on Cummins (1980), "The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue." *TESOL Quarterly*, *50*(4), 940-944. https://doi.org/10.1002/tesq.339

Cummins, J., & Early, M. (Eds.). (2011). *Identity texts: The collaborative creation of power in multilingual schools*. Trentham Books.

Cummins, J., Hu, S., Markus, P., & Kristiina Montero, M. (2015). Identity texts and academic achievement: Connecting the dots in multilingual school contexts. *TESOL Quarterly, 49*(3), 555-581. https://doi.org/10.1002/tesq.241

Dakin, J. (2012). Writing bilingual stories: Developing children's literacy through home languages. In D. Mallows (Ed.), *Innovations in English language teaching for migrants*

and refugees (pp. 13-21). British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/c328_innovations_book_final_2_web.pdf

de la Campa, J. C., & Nassaji, H. (2009). The amount, purpose, and reasons for using L1 in L2 classrooms. *Foreign Language Annals*, *42*(4), 742-759. https://doi.org/10.1111/j.1944-9720.2009.01052.x

de la Colina, A., & del Pilar García Mayo, M. (2009). Oral interaction in task-based EFL learning: The use of the L1 as a cognitive tool. *IRAL - International Review of Applied Linguistics in Language Teaching*, *47*(3-4), 325-345. https://doi.org/10.1515/iral.2009.014

Dickinson, D. K., McCabe, A., Clark–Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics*, *25*(3), 323-347. https://doi.org/10.1017/s0142716404001158

Dicks, J., & Genesee, F. (2016). Bilingual education in Canada. In O. García et al. (Eds.), *Bilingual and multilingual education* (pp. 1-15). Springer International Publishing. https://doi.org/10.1007/978-3-319-02324-3_32-1

Dickson, R., Cherry, M. G., & Boland, A. (2017). Carrying out a systematic review as a Master's thesis. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 1-19). SAGE Publications.

Domínguez de Ramírez, R. D., & Shapiro, E. S. (2007). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English language learners in bilingual education classrooms. *Psychology in the Schools*, *44*(8), 795-806. https://doi.org/10.1002/pits.20266

Dundar, Y., & Fleeman, N. (2017a). Developing my search strategy. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 61-78). SAGE Publications.

Dundar, Y., & Fleeman, N. (2017b). Applying inclusion and exclusion criteria. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 79-91). SAGE Publications.

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology*, *85*(3), 453-465. https://doi.org/10.1037/0022-0663.85.3.453

Edelsky, C., Hudelson, S., Flores, B., Barkin, F., Altwerger, B., & Jilbert, K. (1983). Semilingualism and language deficit. *Applied linguistics*, *4*(1), 1-22.

Ekstrand, L. H. (1977). *Social and individual frame factors in L2 learning: Comparative aspects* [Conference presentation]. Scandinavian Conference on Bilingualism, Helsinki, Finland.

EPPI-Centre. (2019). *EPPI-Centre database of education research.* EPPI-Centre. Retrieved January 28, 2023, from http://eppi.ioe.ac.uk/webdatabases/Intro.aspx?ID=6.

Ervin-Tripp, S. M. (1974). Is second language learning like the first. *TESOL Quarterly*, *8*(2), 111-127. https://doi.org/10.2307/3585535

Espinosa, C., Ascenzi-Moreno, L., & Vogel, S. (2016). *A translanguaging pedagogy for writing: A CUNY-NYSIEB guide for educators*. CUNY-NYSIEB.

Fathman, A. (1975). The relationship between age and second language productive ability. *Language Learning*, *25*(2), 245-253. https://doi.org/10.1111/j.1467-1770.1975.tb00244.x

Fleeman, N., & Dundar, Y. (2017). Data extraction: Where do I begin? In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 93-106). SAGE Publications.

García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Wiley Blackwell.

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.

Gebauer, S. K., Zaunbauer, A. C. M., & Möller, J. (2013). Cross-language transfer in English immersion programs in Germany: Reading comprehension and reading fluency. *Contemporary Educational Psychology*, *38*(1), 64-74. https://doi.org/10.1016/j.cedpsych.2012.09.002

Genesee, F. (1979). Acquisition of reading skills in immersion programs. *Foreign Language Annals*, *12*(1), 71-77. https://doi.org/10.1111/j.1944-9720.1979.tb00952.x

Genesee, F., & Hamayan, E. (1980). Individual differences in second language learning. *Applied Psycholinguistics*, *1*(1), 95-110. https://doi.org/10.1017/s0142716400000758

Geva, E., Wade-Woolley, L., & Shany, M. (1997). Development of reading efficiency in first and second language. *Scientific Studies of Reading*, *1*(2), 119-144. https://doi.org/10.1207/s1532799xssr0102_2

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

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationships between phonological sensitivity, syntactic processing, and verbal working memory in the reading performance of third-grade children. *Journal of Experimental Child Psychology*, *63*(3), 563-582. https://doi.org/10.1006/jecp.1996.0062

Greene, J. P. (1997). A meta-analysis of the Rossell and Baker review of bilingual education research. *Bilingual Research Journal*, *21*(2-3), 103-122. https://doi.org/10.1080/15235882.1997.10668656

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3-15. https://doi.org/10.1016/0093-934x(89)90048-5

Halliday, M. A. K. (1975). *Learning to mean: Explorations in the development of language*.

Edward Arnold.

Hansegård, N. E. (1968). *Tvåspråkighet eller halvspråkighet?* Aldus/Bonniers.

Harbord, J. (1992). The use of the mother tongue in the classroom. *ELT Journal*, *46*(4), 350-355. https://doi.org/10.1093/elt/46.4.350

Hawkins, E. (1987). *Modern languages in the curriculum*. Cambridge University Press.

Hemàndez-Chávez, E., Burt, M., & Dulay, H. (1978). Language dominance and proficiency testing: Some general considerations. *NABE Journal*, *3*(1), 41-54. https://doi.org/10.1080/08855072.1978.10668343.

Hesson, S., Seltzer, K., & Woodley, H. H. (2014). *Translanguaging in curriculum and instruction: A CUNY-NYSIEB guide for educators*. CUNY-NYSIEB.

Howatt, A. P. R., & Widdowson, H. G. (2004). *A history of ELT*. Oxford University Press.

Huennekens, M. E. (2013). *The cross-linguistic effects of dialogic reading on young English language learners*. [Unpublished doctoral dissertation, Virginia Commonwealth University].

Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of Experimental Child Psychology*, *82*(1), 2-28. https://doi.org/10.1006/jecp.2002.2670

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258-286). Cambridge University Press. http://doi.org/10.1017/CBO9781139524780.011

IDESR. (n.d.). *International Database of Education Systematic Reviews*. IDESR. Retrieved January 28, 2023, from https://idesr.org

Karlander, D., & Salö, L. (2023). The origin of semilingualism: Nils-Erik Hansegård and the cult

of the mother tongue. *Journal of Sociolinguistics,* 1-20. https://doi.org/10.1111/josl.12614

Kenner, C., Al-Azami, S., Gregory, E., & Ruby, M. (2008). Bilingual poetry: Expanding the cognitive and cultural dimensions of children's learning. *Literacy*, *42*(2), 92-100. https://doi.org/10.1111/j.1741-4369.2008.00486.x

Keung, Y.-C., & Ho, C. S.-H. (2009). Transfer of reading-related cognitive skills in learning to read Chinese (L1) and English (L2) among Chinese elementary school children. *Contemporary Educational Psychology*, *34*(2), 103-112. https://doi.org/10.1016/j.cedpsych.2008.11.001

Kobayashi, M. (2003). The role of peer support in ESL students' accomplishment of oral academic tasks. *Canadian Modern Language Review*, *59*(3), 337-369. https://doi.org/10.3138/cmlr.59.3.337

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

Krashen, S., & McField, G. (2005). What works? Reviewing the latest evidence on bilingual education. *Language Learner*, *1*(2), 7-10.

Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press.

Lambert, W. E., & Tucker, G. R. (1972). *Bilingual education of children: The St. Lambert experiment*. Newbury House.

Lambert, W. E., Tucker, G. R., & d'Anglejan, A. (1973). Cognitive and attitudinal consequences of bilingual schooling. *Journal of Educational Psychology*, *65*(2), 141-159. https://doi.org/10.1037/h0034983.

Lapkin, S., & Swain, M. (1977). The use of English and French cloze tests in a bilingual education program evaluation: Validity and error analysis. *Language Learning*, *27*(2), 279-310. https://doi.org/10.1111/j.1467-1770.1977.tb00123.x

Larsen-Freeman, D. (2011). Complex, dynamic systems: A new transdisciplinary theme for

applied linguistics? *Language Teaching*, *45*(2), 202-214. https://doi.org/10.1017/s0261444811000061

Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal*, *97*(4), 887-901. https://doi.org/10.1111/j.1540-4781.2013.12044.x

Littlewood, W., & Yu, B. (2011). First language and target language in the foreign language classroom. *Language Teaching*, *44*(1), 64-77. https://doi.org/10.1017/s0261444809990310

López, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of Hispanic head start children. *Bilingual Research Journal*, *28*(1), 1-18. https://doi.org/10.1080/15235882.2004.10162609

Lugo-Neris, M. J., Jackson, C. W., & Goldstein, H. (2010). Facilitating vocabulary acquisition of young English language learners. *Language, Speech, and Hearing Services in Schools*, *41*(3), 314-327. https://doi.org/10.1044/0161-1461(2009/07-0082)

Macaro, E. (2019). Systematic reviews in applied linguistics. In J. Mckinley, & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 230-239). Routledge. https://doi.org/10.4324/9780367824471-20

Macnamara, J. (1966). *Bilingualism and primary education: A study of Irish experience*. Edinburgh University Press.

MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contributions to a deficit view of linguistic minorities. *Hispanic Journal of Behavioral Sciences*, *22*(1), 3-45. https://doi.org/10.1177/0739986300221001

Martin-Jones, M., & Romaine, S. (1986). Semilingualism: A half-baked theory of communicative competence. *Applied linguistics*, *7*(1), 26-38.

May, S. (2014). *The multilingual turn: Implications for SLA, TESOL and bilingual education*. Routledge.

McBride-Chang, C. (1995). What is phonological awareness? *Journal of Educational Psychology*, *87*(2), 179-192. https://doi.org/10.1037/0022-0663.87.2.179

McField, G. (2002). *Does program quality matter? A meta-analysis of select bilingual education studies* (Publication No. 3093791) [Doctoral dissertation, University of Southern California]. ProQuest Dissertations Publishing.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*(1), 114-135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Mitchell, R. & Myles, F. (2001). Second language learning: Key concepts and issues. In C. N. Candlin & N. Mercer (Eds.), *English language teaching in its social context: A reader* (pp. 11-27). Routledge.

Moher, D., Pham, B., Lawson, M., & Klassen, T. (2003). The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment*, *7*(41). https://doi.org/10.3310/hta7410

Murphy, V. A., Macaro, E., Alba, S., & Cipolla, C. (2014). The influence of learning a second language in primary school on developing first language literacy skills. *Applied Psycholinguistics*, *36*(5), 1133-1153. https://doi.org/10.1017/s0142716414000095

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/9781009093873

Nation, I. S. P. (2003). The role of the first language in foreign language learning. *Asian EFL Journal, 5*(2), 1-8.

Oller, J. (1976). A program for language testing research. *Language Learning*, *4*(1), 141-165.

Oller, J. (1977). Attitude variables in second language learning. In M. Burt, H. Dulay, & M. Finocchiaro (Eds.), *Viewpoints on English as a second language* (pp. 172-184). Regents.

Oller, J. (1978). The language factor in the evaluation of bilingual education. In J. Alatis (Ed.), *Georgetown University round table on languages and learning 1978: International dimensions of bilingual education* (pp. 410-422). Georgetown University Press.

Oller, J. & Perkins, K. (1978). *Language in education: Testing the tests*. Newbury.

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, *6*(3). 281-307. https://doi.org/10.1515/applirev-2015-0014

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McGuinness, L. A. (2020). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Parke, T., Drury, R., Kenner, C., & Robertson, L. H. (2002). Revealing invisible worlds: Connecting the mainstream with bilingual children's home and community learning. *Journal of Early Childhood Literacy*, *2*(2), 195-220. https://doi.org/10.1177/14687984020022004

Paulsrud, B., Rosén, J., Boglárka Straszer, & Wedin, Å. (Eds.). (2017). *New perspectives on translanguaging and education*. Multilingual Matters. https://doi.org/10.21832/paulsr7814

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

Preevo, M., Malda, M., Mesman, J., & van IJzendoorn, M. (2016). Within- and cross-language relations between oral language proficiency and school outcomes in bilingual

children with an immigrant background: A meta-analytical study. *Review of Educational Research*, 86(1), 237-276. https://doi.org/10.3102/0034654315584685

Quiroga, T., Lemos-Britton, Z., Mostafapour, E., Abbott, R. D., & Berninger, V. W. (2002). Phonological awareness and beginning reading in Spanish-speaking ESL first graders. *Journal of School Psychology*, *40*(1), 85-111. https://doi.org/10.1016/s0022-4405(01)00095-4

Ramírez, M. C. (2001). An investigation of English language learners and reading skills on reading comprehension for Spanish-speaking English language learners (Publication No. 3024526) [Doctoral dissertation, University of Oregon]. ProQuest Dissertations Publishing.

Ramsey, C. A., & Wright, E. N. (1974). Age and second language learning. *The Journal of Social Psychology*, *94*(1), 115-121. https://doi.org/10.1080/00224545.1974.9923189

Read, J. A. S. (2000). *Assessing vocabulary*. Cambridge University Press.

Reljić, G., Ferring, D., & Martin, R. (2015). A meta-analysis on the effectiveness of bilingual programs in Europe. *Review of Educational Research*, *85*(1), 92-128. https://doi.org/10.3102/0034654314548514

Rolstad, K., Mahoney, K., & Glass, G. V. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*, *19*(4), 572-594. https://doi.org/10.1177/0895904805278067

Sánchez, M. J. (2004). Effect of instruction with expert patterns on the lexical learning of English as a foreign language. *System*, *32*(1), 89-102. https://doi.org/10.1016/j.system.2003.05.002

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language proficiency divisible or unitary? A factor analysis of 22 English language proficiency tests. In J. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 13-23). Newbury House.

Scott, V. M., & de la Fuente, M. J. (2008). What's the problem? L2 learners' use of the L1 during consciousness-raising, form-focused tasks. *The Modern Language Journal*, *92*(1), 100-113. https://doi.org/10.1111/j.1540-4781.2008.00689.x

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*(1), 747-770.

Sieh, Y.-C. (2008). A possible role for the first language in young learners' processing and storage of foreign language vocabulary. *Annual Review of Education, Communication, and Language Sciences, 5*(1), 136-160.

Skutnabb-Kangas, T., & Toukomaa, P. (1976). *Teaching migrant children's mother tongue and learning the language of the host country in the context of the socio-cultural situation of the migrant family*. The Finnish National Commission for UNESCO.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5-11. https://doi.org/10.3102/0013189x015009005

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research*, *75*(2), 247-284. https://doi.org/10.3102/00346543075002247

Snow, C. E., & Hoefnagel-Hohle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, *49*(4), 1114-1128. https://doi.org/10.2307/1128751

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology, 100*(1), 162-174.

https://doi.org/10.1037/0022-0663.100.1.162

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*(1S), 282S-306S. https://doi.org/10.3102/0002831216634463

Storch, N., & Wigglesworth, G. (2003). Is there a role for the use of the L1 in an L2 setting? *TESOL Quarterly*, *37*(4), 760-770. https://doi.org/10.2307/3588224

Swain, M., & Lapkin, S. (2000). Task-based second language learning: The uses of the first language. *Language Teaching Research*, *4*(3), 251-274. https://doi.org/10.1177/136216880000400304

Swain, M., Lapkin, S., & Barik, H. C. (1976). The cloze test as a measure of second language proficiency for young children. *Working Papers on Bilingualism 11*, 32-42.

Taft, R., & Bodi, M. (1980). A study of language competence and first language maintenance in bilingual children. *Applied Psychology*, *29*(1-2), 173-182. https://doi.org/10.1111/j.1464-0597.1980.tb00889.x

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing*, *1*(3), 176-184. https://doi.org/10.1111/j.1524-475x.2004.04006.x

Thomas, J., O'Mara-Eves, A., Harden, A., & Newman, M. (2017). Synthesis methods for combining and configuring textual or mixed methods data. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 181-209). SAGE Publications.

Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, *59*(3), 623-646. https://doi.org/10.1111/j.1467-9922.2009.00519.x

Toukomaa, P., & Skutnabb-Kangas, T. (1977). *The intensive teaching of the mother tongue to migrant children of pre-school age and children in the lower level of comprehensive school*. The Finnish National Commission for UNESCO.

UNESCO (1953). The Use of Vernacular Language in Education: Report of UNESCO Meeting of Specialists.

van Gelderen, A., Schoonen, R., Stoel, R. D., de Glopper, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, *99*(3), 477-491. https://doi.org/10.1037/0022-0663.99.3.477

van Weijen, D., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing*, *18*(4), 235-250. https://doi.org/10.1016/j.jslw.2009.06.003

Walters, K., & Gunderson, L. (1985). Effects of parent volunteers reading first language (L1) books to ESL students. *Reading Teacher, 39*(1), 66-69.

Wang, M., Cheng, C., & Chen, S.-W. (2006). Contribution of morphological awareness to Chinese-English biliteracy acquisition. *Journal of Educational Psychology*, *98*(3), 542-553. https://doi.org/10.1037/0022-0663.98.3.542

Wang, M., Park, Y., & Lee, K. R. (2006). Korean-English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Educational Psychology*, *98*(1), 148-158. https://doi.org/10.1037/0022-0663.98.1.148

Wang, M., Perfetti, C. A., & Liu, Y. (2005). Chinese-English biliteracy acquisition: Cross-language and writing system transfer. *Cognition*, *97*(1), 67-88. https://doi.org/10.1016/j.cognition.2004.10.001

Wang, M., Yang, C., & Cheng, C. (2009). The contributions of phonology, orthography, and morphology in Chinese-English biliteracy acquisition. *Applied Psycholinguistics*, *30*(2), 291-314. https://doi.org/10.1017/s0142716409090122

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715-738. https://doi.org/10.1111/modl.12671

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching, 56*(2), 161-180. https://doi.org/10.1017/s0261444822000507

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9-30. https://doi.org/10.1093/applin/amx039

Williams, C. (1994). *Arfarniad o Ddulliau Dysgu ac Addysgu yng Nghyd-destun Addysg Uwchradd Ddwyieithog* (Publication No. 10310435) [Doctoral dissertation, Bangor University]. ProQuest Dissertations Publishing.

Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research*, *55*(3), 269-317.

Yiakoumetti, A. (2006). A bidialectal programme for the learning of standard modern Greek in Cyprus. *Applied Linguistics*, *27*(2), 295-317. https://doi.org/10.1093/applin/aml012

Appendices

Appendix 1: The PRISMA Checklist by Page et al. (2020) 78

Appendix 2: The Protocol Document 80

Appendix 3: The Three Fields of Terms Used for the Electronic Searches 86

Appendix 4: The Five Boolean Strings Used for the Electronic Searches 88

Appendix 5a: A Screenshot of Rayyan Showing Double Screening 93

Appendix 5b: A Screenshot of Rayyan Showing Two Conflicting Decisions 93

Appendix 6: The Eligibility Outcomes of Full-Text Screening 94

Appendix 7: An Example Data Extraction Form 96

Appendix 8: The Adapted EPHPP Risk of Bias Assessment Tool 100

Appendix 9: The Reference List of Included Reports 104

Appendix 10: A Summary of Key Descriptive Data from the Included Studies 108

Appendix 11: The EPHPP Risk of Bias Assessment Tool Guidelines 121

Appendix 12: Study Outcomes in Relation to Intervention Type and Quality Assessment 125

Appendix 1: The PRISMA Checklist by Page et al. (2020)

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Cover page
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page i
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Pages 1-20
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 21
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Pages 22-24
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 24-25
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 25
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 26
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 27
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 27
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Page 27
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Pages 27-28
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	N/A
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 29
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Page 29
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Page 29
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Page 29
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	N/A
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	N/A
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Pages 27-28
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	N/A

Section and Topic	Item #	Checklist item	Location where item is reported
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Pages 29-30
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Pages 29-30
Study characteristics	17	Cite each included study and present its characteristics.	Pages 31-36
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Pages 36-41
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	N/A
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	N/A
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	N/A
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	N/A
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	N/A
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Pages 42-55
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Pages 42-55
	23b	Discuss any limitations of the evidence included in the review.	Page 56
	23c	Discuss any limitations of the review processes used.	Page 56
	23d	Discuss implications of the results for practice, policy, and future research.	Page 57
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Appendix 2
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Appendix 2
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	N/A
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	N/A
Competing interests	26	Declare any competing interests of review authors.	N/A
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Appendices

Appendix 2: The Protocol Document

Protocol Registration Form

The content of this form is based on Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1). doi: 10.1186/2046-4053-4-1. We strongly recommended that this form is completed in accordance with the recommendations in that paper.

Title

The effects of using the first language in comparison to the second language to teach target language vocabulary items in primary or secondary school settings: Protocol for a systematic review.

The title should identify the report as a protocol of a systematic review. If this is an update of an existing review, indicate this in the title. Examples: The effectiveness of task-based language teaching for adult learners of foreign or second languages. Protocol for a systematic review. *Or* The effectiveness of written corrective feedback for the acquisition of L2 grammar among primary-aged language learners. A protocol for an update of the systematic review by Bloggs and Smith (2020).

Main Contact/Corresponding Author

Name: Lucy Alice Armstrong

Institutional affiliation: University of Oxford

e-mail address: lucy.armstrong@education.ox.ac.uk

Physical mailing address: Department of Education, 15 Northam Gardens, Oxford, OX2 6PY

Additional Authors

Name: Hamish Chalmers

Institutional affiliation: University of Oxford

e-mail address: hamish.chalmers@education.ox.ac.uk

On the IDESR website, you will be able to add as many additional authors as you need.

Review Question(s)

What are the effects of using the first language in comparison to the second language to teach target language vocabulary items in primary or secondary school settings?

Provide the review questions. For reviews of interventions, include reference to PI(E)CO, as appropriate (Participants, Intervention (Exposure), Comparator, Outcomes). Examples: What are the effects of study abroad compared to classroom teaching on vocabulary acquisition among adolescent learners of a foreign language? *Or* What is the impact on academic attainment of attending a bilingual school compared to a target language only school among minority language users?

Rationale

Following the multilingual turn in language teaching, and the recent proliferation of pedagogical approaches that advise teachers to utilise the L1 of their learners in L2 lessons, this review aims to establish the nature and extent of empirical research in support of these claims. It will do so with regard to vocabulary teaching in particular because a previous systematic review into the effects of L1-mediated teaching in L2 settings on L2 proficiency (Chalmers, 2019) found only ten studies, and five of these examined L1-mediated approaches to L2 vocabulary acquisition. This provided insight into the scope of research on this topic at the primary level, but not beyond that.

In no more than 300 words, describe the rationale for the review in the context of what is already known.

Inclusion Criteria

Bibliographic information

Include 1: Studies with either a complete reference, or sufficient bibliographic information to enable the location of their full text.
Exclude 1: Studies with an incomplete reference, or insufficient bibliographic information.
Rationale: Without sufficient bibliographic information, such as the full title or list of authors, attempting to locate a study is likely to be time-consuming and/or futile.

Focus of the systematic review

Include 2: Relevant to assessing the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition.
Exclude 2: Not relevant to assessing the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition.
Rationale: This systematic review intends to contrast the effects of using either the L1 or the L2 to teach L2 vocabulary items, and so studies that do not directly compare the two are irrelevant.

Publication date

Include 3: Published between 1980 and 2022.
Exclude 3: Published before 1980 or after 2022.
Rationale: (1) Cummins posited his Linguistic Interdependence Hypothesis, which speaks to the role that the L1 may play in L2 acquisition, in 1979. His distinction between Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP) then followed in 1980, along with his Common Underlying Proficiency Hypothesis, and so research responding to these theoretical ideas will be from after 1980. (2) A previous systematic review by Chalmers (2019) examined the effects of using L1-mediated teaching approaches in L2 pre-primary and primary school settings, and focused on the years 1980 to 2015. Since this review aims to build on his findings by examining the effects of L1-mediated teaching approaches on vocabulary acquisition in particular, a similar starting date with an expanded range seems appropriate.

Study design

Include 4: Intervention studies in the form of randomised control trials or quasi-experiments.
Exclude 4: Non-intervention studies.
Rationale: This review aims to determine whether teachers should be advised to use L1-mediated approaches or L2-mediated approaches when teaching L2 vocabulary, and so it is necessary to examine empirical evidence.

Study setting
Include 5: Studies of formal education, focusing on typical school/student populations.
Exclude 5: Studies of informal education, such as private language centres, or studies that focus on non-typical school/student populations, such as those with special educational needs.
Rationale: Students with special educational needs are likely to have specific requirements that set them apart from mainstream learners, and confound research into the effects of L1-mediated teaching approaches and L2-mediated teaching approaches on L2 vocabulary acquisition for non-special needs learners by producing quite different results.

Study participants
Include 6: Conducted with primary and/or secondary school-aged children.
Exclude 6: Conducted with pre-school-aged children, university students, or adults.
Rationale: The systematic review that Chalmers (2019) conducted was technically a scoping review, designed to examine the nature and extent of empirical research into the effects of L1-mediated teaching approaches in L2 educational settings in general. As it found some evidence to suggest that using the L1 to teach L2 vocabulary is beneficial at the primary level, however, it seems sensible to ascertain whether (a) further research has been undertaken on this topic, and (b) this holds true at the secondary level, as well.

Note: Given that the age of students at different levels of education varies slightly between countries, the definition of what constitutes primary or secondary school-aged students in the context of each study will determine its inclusion, rather than the age of the learners.

Study outcomes
Include 7: Outcomes are quantitative measures of L2 vocabulary acquisition, either in the form of target L2 word knowledge, or the comprehension of texts containing target L2 words.
Exclude 7: Outcomes are qualitative measures of L2 vocabulary acquisition, or quantitative measures that are either self-reported, or assess an outcome other than target L2 word knowledge, or the comprehension of texts containing target L2 words.
Rationale: Although qualitative measures of vocabulary acquisition could help to shed light on how students feel about L1-mediated and L2-mediated vocabulary teaching approaches – and therefore whether they would be well-received in schools – the focus of this review is on whether L1-mediated or L2-mediated approaches facilitate greater vocabulary acquisition.

Type of publication
Include 8: Any type of publication, with grey literature included.
Exclude 8: No type of publication will be excluded.
Rationale: This systematic review aims to synthesise all research into the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition over a 43-year period, regardless of the form that research takes.

List here the criteria to be used for inclusion in the review. As appropriate, include information about the population, interventions, comparators, primary outcomes, setting, study design(s), time frame, publication types, language(s) of publication, etc.

Information Sources

In order for this review to be as comprehensive as possible, the following databases in the fields of education, applied linguistics, psychology and/or a range of disciplines will be consulted:

- ProQuest Social Science Premium Collection
- British Education Index
- Web of Science
- PsychINFO
- SCOPUS

They will be accessed electronically through subscription to the University of Oxford's Bodleian Library, and the ProQuest Dissertation & Theses Global, EThOS and OpenGrey databases will also be searched in order to look specifically for any additional grey literature.

Describe all intended information sources. Include the names of electronic databases, journals or websites that will be hand searched, contact with study authors, grey literature sources, etc.

Search Strategy

With the support of both my dissertation supervisor and a subject consultant at the Bodleian Education Library, five Boolean strings will be developed for use in the five databases outlined above. In total, three fields of terms have already been identified as essential for use in these strings – those relating to the notions of a first language, second language and vocabulary – and once extensive pilot searches have been conducted with a range of potential terms within each field, those that are the most appropriate will be selected. After this, they will then be developed into Boolean strings using the different field codes of each bibliographic database, but in general, the three fields will be united by 'AND' functions, and their individual terms by 'OR' functions.

Present the search strategy to be used for at least one electronic database such that it could be repeated by a third party. Include planned limiters, for example, date range, and location in the text (e.g. Title, Abstract, or Full Text). Present these as a Boolean phrase if possible. If Boolean phrasing is inappropriate for your review, present the search strategy in a way that can allow replication by a third party.

Data Management

Following the application of five Boolean search strings to each of the databases identified, the bibliographic information and abstracts of relevant results will be exported from those databases in .RIS format, and uploaded to the Rayyan systematic review management software.

Describe the mechanisms by which the data will be managed throughout the review. For example, say which data management software the review team use, e.g. Rayyan, EPPI Reviewer, Excel, Covidence, etc. Describe if these will change with different phases of the review (abstract screening, full text screening, data extraction, etc.).

Selection Process

Once uploaded to Rayyan, duplicates of records will be removed, and those that remain will be screened for inclusion or exclusion against the eligibility criteria of the review. When excluded, justification for this decision will be documented using 'Exclude 1' or 'Exclude 2' labels, etc., and if a record cannot be excluded on the basis of the information in its title or abstract, then it will be marked as 'Include' until full-text screening, at which point a final decision will be made.

Describe the method by which studies will be selected for inclusion at each stage of the review. For example, how many reviewers will screen abstracts/full texts? What quality assurance procedures will be in place in each of these phases (dual screening of all records, percentage dual screened then checked for consistency, etc.)?

Data Collection Process

When all of the studies that meet the inclusion criteria for this systematic review have finally been selected, they will be read in detail, and all pertinent data collected by means of a bespoke data extraction form that will be developed with reference both to the examples provided by Petticrew and Roberts (2006), and that utilised by Chalmers (2019) in his systematic review.

Describe how data will be extracted from reports. Will a data extraction form be used? Will this be piloted? Will data be extracted independently by multiple reviewers? What is the process for obtaining data not contained in the reports (e.g. contacting authors directly)?

Data Items

Although it has not yet been designed, the intention is for the data extraction form to compile both descriptive and analytical data, and to do so with regard to the four categories of:

(1) administrative information
(2) contextual information
(3) intervention information
(4) outcome information

List and define all data items that will be extracted (e.g. participant info, outcome measures, sources of funding, study design, etc.).

Risk of bias/trustworthiness of individual studies

After data extraction, the trustworthiness of each included study will be assessed using an adapted version of the Quality Assessment Tool for Quantitative Studies by the Effective Public Health Practice Project (EPHPP), which calculates both an overall 'weak', 'moderate' or 'strong' rating for a study, and its quality according to six methodological components.

Describe how risk of bias, trustworthiness, or quality of individual studies will be assessed. Name any specific tools, e.g. Gorard's Sieve, Maryland Scientific Methods Scale, Cochrane Risk of Bias Tool, EPPI Weight of Evidence Tool, etc. State how this information will be used in the synthesis.

Data Synthesis

If amendable to statistical synthesis, then the data extracted from the included studies will be meta-analysed, but given the heterogeneity of the research comparing L1-mediated vocabulary teaching approaches with L2-mediated vocabulary teaching approaches at the primary level that Chalmers (2019) uncovered, this seems unlikely. As such, a narrative approach to synthesis will instead be adopted, and the findings of the review categorised by their principal themes.

Describe criteria under which quantitative synthesis will be performed. If quantitative synthesis is appropriate, describe preferred summary measure (Cohen's D, Hedges' G, etc.) and how these will be combined. Describe any additional planned analyses (e.g. sub-group analysis). If quantitative synthesis is not appropriate, describe how data will be combined and summarised.

Meta-biases

While the problem of publication bias will not be directly addressed in this review, it will be acknowledged as undermining the accuracy of any conclusions derived from the accumulated empirical data, and as suggesting that the effects of various L1-mediated or L2-mediated vocabulary teaching interventions may be overstated.

Describe how meta-biases (publication bias, selective outcome reporting, etc.) will be addressed.

Confidence in cumulative evidence

How confidence in cumulative evidence will be calculated has not yet been determined.

Describe how the strength of the body of evidence will be assessed.

Sources of Funding

This review has no sources of funding.

Specify any financial or other support for the review. Provide names for the review funder and/or sponsor. Include the funding reference number if available.

Role of Funders

N/A.

Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol.

Anticipated or actual start date: March 2023

Anticipated completion date: August 2023

Other language resources

N/A.

If you have other language versions of your protocol, please include a link here.

Current Status

☒ Ongoing
☐ Completed but not published
☐ Completed and published

For initial submissions select 'ongoing'. When you have completed the review, you should return to update the record accordingly. Your review status will be displayed next to your record in IDESR.

Details of Published Review

N/A.

Once the review is complete, you should return to this form to add information about where it has been published. Include the full bibliographic reference and a link to the published document, a DOI or URL. You should link to a pre-print as soon as one is available, then update this information once the full review has been published. Your review will then be added to the IDESR database.

Appendix 3: The Three Fields of Terms Used for the Electronic Searches

Field 1	AND	**Field 2**	AND	**Field 3**
L1	“	L2	“	"vocabulary acquisition"
"first language*"	“	"second language*"	“	"vocabulary attainment"
"home language*"	“	"additional language*"	“	"vocabulary gain"
"native language*"	“	"non-native language*"	“	"vocabulary growth"
"mother tongue*"	“	"foreign language*"	“	"vocabulary knowledge"
"heritage language*"	“	"majority language*"	“	"vocabulary learning"
"community language*"	“	"school language*"	“	"vocabulary retention"
"background language*"	“	"target language*"	“	"vocabulary increase"
"family language*"	“	"academic language*"	“	"vocabulary improvement"
"ethnic language*"	“	"mainstream language*"	“	"vocabulary development"
"minority language*"	“	"education* language*"	“	"vocabulary expansion"
translanguaging	“	"instructional language*"	“	"vocabulary recall"
codeswitching	“	"language of education"	“	"word acquisition"
"code-switching"	“	"language of instruction"	“	"word attainment"
"code switching"	“	MFL	“	"word gain"
codemixing	“	ESL	“	"word growth"
"code-mixing"	“	EAL	“	"word knowledge"
"code mixing"	“	EFL	“	"word learning"
codemeshing	“	"English for speakers of other languages"	“	"word retention"
"code-meshing"	“	ESOL	“	"word increase"
"code meshing"	“	"English language teaching"	“	"word improvement"
		ELT	“	"word development"
		TESL	“	"word expansion"
		TEAL	“	"word recall"
		TEFL	“	"lex* acquisition"
		TESOL	“	"lex* attainment"
		"teaching English to speakers of other languages"	“	"lex* gain"
				"lex* growth"
				"lex* knowledge"
				"lex* learning"
				"lex* retention"
				"lex* increase"
				"lex* improvement"
				"lex* development"
				"lex* expansion"
				"lex* recall"
				"acquisition of vocabulary"
				"attainment of vocabulary"
				"gain in vocabulary"
				"growth in vocabulary"
				"knowledge of vocabulary"
				"learning of vocabulary"

				"retention of vocabulary"
				"increase in vocabulary"
				"improvement in vocabulary"
				"development of vocabulary"
				"expansion of vocabulary"
				"recall of vocabulary"
				"acquisition of word*"
				"attainment of word*"
				"gain in word*"
				"growth in word*"
				"knowledge of word*"
				"learning of word*"
				"retention of word*"
				"increase in word*"
				"improvement in word*"
				"development of word*"
				"expansion of word*"
				"recall of word*"
				"acquisition of lex*"
				"attainment of lex*"
				"gain in lex*"
				"growth in lex*"
				"knowledge of lex*"
				"learning of lex*"
				"retention of lex*"
				"increase in lex*"
				"improvement in lex*"
				"development of lex*"
				"expansion of lex*"
				"recall of lex*"

Appendix 4: The Five Boolean Strings Used for the Electronic Searches

Database 1: The ProQuest Social Science Premium Collection

(NOFT(L1) OR NOFT("first language*") OR NOFT("home language*") OR NOFT("native language*") OR NOFT("mother tongue*") OR NOFT("heritage language*") OR NOFT("community language*") OR NOFT("background language*") OR NOFT("family language*") OR NOFT("ethnic language*") OR NOFT("minority language*") OR NOFT(translanguaging) OR NOFT(codeswitching) OR NOFT("code-switching") OR NOFT("code switching") OR NOFT(codemixing) OR NOFT("code-mixing") OR NOFT("code mixing") OR NOFT(codemeshing) OR NOFT("code-meshing") OR NOFT("code meshing")) AND (NOFT(L2) OR NOFT("second language*") OR NOFT("additional language*") OR NOFT("non-native language*") OR NOFT("foreign language*") OR NOFT("majority language*") OR NOFT("school language*") OR NOFT("target language*") OR NOFT("academic language*") OR NOFT("mainstream language*") OR NOFT("education* language*") OR NOFT("instructional language*") OR NOFT("language of education") OR NOFT("language of instruction") OR NOFT(MFL) OR NOFT(ESL) OR NOFT(EAL) OR NOFT(EFL) OR NOFT("English for speakers of other languages") OR NOFT(ESOL) OR NOFT("English language teaching") OR NOFT(ELT) OR NOFT(TESL) OR NOFT(TEAL) OR NOFT(TEFL) OR NOFT(TESOL) OR NOFT("teaching English to speakers of other languages")) AND (NOFT("vocabulary acquisition") OR NOFT("vocabulary attainment") OR NOFT("vocabulary gain") OR NOFT("vocabulary growth") OR NOFT("vocabulary knowledge") OR NOFT("vocabulary learning") OR NOFT("vocabulary retention") OR NOFT("vocabulary increase") OR NOFT("vocabulary improvement") OR NOFT("vocabulary development") OR NOFT("vocabulary expansion") OR NOFT("vocabulary recall") OR NOFT("word acquisition") OR NOFT("word attainment") OR NOFT("word gain") OR NOFT("word growth") OR NOFT("word knowledge") OR NOFT("word learning") OR NOFT("word retention") OR NOFT("word increase") OR NOFT("word improvement") OR NOFT("word development") OR NOFT("word expansion") OR NOFT("word recall") OR NOFT("lex* acquisition") OR NOFT("lex* attainment") OR NOFT("lex* gain") OR NOFT("lex* growth") OR NOFT("lex* knowledge") OR NOFT("lex* learning") OR NOFT("lex* retention") OR NOFT("lex* increase") OR NOFT("lex* improvement") OR NOFT("lex* development") OR NOFT("lex* expansion") OR NOFT("lex* recall") OR NOFT("acquisition of vocabulary") OR NOFT("attainment of vocabulary") OR NOFT("gain in vocabulary") OR NOFT("growth in vocabulary") OR NOFT("knowledge of vocabulary") OR NOFT("learning of vocabulary") OR NOFT("retention of vocabulary") OR NOFT("increase in vocabulary") OR NOFT("improvement in vocabulary") OR NOFT("development of vocabulary") OR NOFT("expansion of vocabulary") OR NOFT("recall of vocabulary") OR NOFT("acquisition of word*") OR NOFT("attainment of word*") OR NOFT("gain in word*") OR NOFT("growth in word*") OR NOFT("knowledge of word*") OR NOFT("learning of word*") OR NOFT("retention of word*") OR NOFT("increase in word*") OR NOFT("improvement in word*") OR NOFT("development of word*") OR NOFT("expansion of word*") OR NOFT("recall of word*") OR NOFT("acquisition of lex*") OR NOFT("attainment of lex*") OR NOFT("gain in lex*") OR NOFT("growth in lex*") OR NOFT("knowledge of lex*") OR NOFT("learning of lex*") OR NOFT("retention of lex*") OR NOFT("increase in lex*") OR NOFT("improvement in lex*") OR NOFT("development of lex*") OR NOFT("expansion of lex*") OR NOFT("recall of lex*"))

Database 2: The British Education Index

(AB(L1) OR AB("first language*") OR AB("home language*") OR AB("native language*") OR AB("mother tongue*") OR AB("heritage language*") OR AB("community language*") OR AB("background language*") OR AB("family language*") OR AB("ethnic language*") OR AB("minority language*") OR AB(translanguaging) OR AB(codeswitching) OR AB("code-switching") OR AB("code switching") OR AB(codemixing) OR AB("code-mixing") OR AB("code mixing") OR AB(codemeshing) OR AB("code-meshing") OR AB("code meshing")) AND (AB(L2) OR AB("second language*") OR AB("additional language*") OR AB("non-native language*") OR AB("foreign language*") OR AB("majority language*") OR AB("school language*") OR AB("target language*") OR AB("academic language*") OR AB("mainstream language*") OR AB("education* language*") OR AB("instructional language*") OR AB("language of education") OR AB("language of instruction") OR AB(MFL) OR AB(ESL) OR AB(EAL) OR AB(EFL) OR AB("English for speakers of other languages") OR AB(ESOL) OR AB("English language teaching") OR AB(ELT) OR AB(TESL) OR AB(TEAL) OR AB(TEFL) OR AB(TESOL) OR AB("teaching English to speakers of other languages")) AND (AB("vocabulary acquisition") OR AB("vocabulary attainment") OR AB("vocabulary gain") OR AB("vocabulary growth") OR AB("vocabulary knowledge") OR AB("vocabulary learning") OR AB("vocabulary retention") OR AB("vocabulary increase") OR AB("vocabulary improvement") OR AB("vocabulary development") OR AB("vocabulary expansion") OR AB("vocabulary recall") OR AB("word acquisition") OR AB("word attainment") OR AB("word gain") OR AB("word growth") OR AB("word knowledge") OR AB("word learning") OR AB("word retention") OR AB("word increase") OR AB("word improvement") OR AB("word development") OR AB("word expansion") OR AB("word recall") OR AB("lex* acquisition") OR AB("lex* attainment") OR AB("lex* gain") OR AB("lex* growth") OR AB("lex* knowledge") OR AB("lex* learning") OR AB("lex* retention") OR AB("lex* increase") OR AB("lex* improvement") OR AB("lex* development") OR AB("lex* expansion") OR AB("lex* recall") OR AB("acquisition of vocabulary") OR AB("attainment of vocabulary") OR AB("gain in vocabulary") OR AB("growth in vocabulary") OR AB("knowledge of vocabulary") OR AB("learning of vocabulary") OR AB("retention of vocabulary") OR AB("increase in vocabulary") OR AB("improvement in vocabulary") OR AB("development of vocabulary") OR AB("expansion of vocabulary") OR AB("recall of vocabulary") OR AB("acquisition of word*") OR AB("attainment of word*") OR AB("gain in word*") OR AB("growth in word*") OR AB("knowledge of word*") OR AB("learning of word*") OR AB("retention of word*") OR AB("increase in word*") OR AB("improvement in word*") OR AB("development of word*") OR AB("expansion of word*") OR AB("recall of word*") OR AB("acquisition of lex*") OR AB("attainment of lex*") OR AB("gain in lex*") OR AB("growth in lex*") OR AB("knowledge of lex*") OR AB("learning of lex*") OR AB("retention of lex*") OR AB("increase in lex*") OR AB("improvement in lex*") OR AB("development of lex*") OR AB("expansion of lex*") OR AB("recall of lex*"))

Database 3: Web of Science Core Collection

(AB=(L1) OR AB=("first language*") OR AB=("home language*") OR AB=("native language*") OR AB=("mother tongue*") OR AB=("heritage language*") OR AB=("community language*") OR AB=("background language*") OR AB=("family language*") OR AB=("ethnic language*") OR AB=("minority language*") OR AB=(translanguaging) OR AB=(codeswitching) OR AB=("code-switching") OR AB=("code switching") OR AB=(codemixing) OR AB=("code-mixing") OR AB=("code mixing") OR AB=(codemeshing) OR AB=("code-meshing") OR AB=("code meshing")) AND (AB=(L2) OR AB=("second language*") OR AB=("additional

language*") OR AB=("non-native language*") OR AB=("foreign language*") OR AB=("majority language*") OR AB=("school language*") OR AB=("target language*") OR AB=("academic language*") OR AB=("mainstream language*") OR AB=("education* language*") OR AB=("instructional language*") OR AB=("language of education") OR AB=("language of instruction") OR AB=(MFL) OR AB=(ESL) OR AB=(EAL) OR AB=(EFL) OR AB=("English for speakers of other languages") OR AB=(ESOL) OR AB=("English language teaching") OR AB=(ELT) OR AB=(TESL) OR AB=(TEAL) OR AB=(TEFL) OR AB=(TESOL) OR AB=("teaching English to speakers of other languages")) AND (AB=("vocabulary acquisition") OR AB=("vocabulary attainment") OR AB=("vocabulary gain") OR AB=("vocabulary growth") OR AB=("vocabulary knowledge") OR AB=("vocabulary learning") OR AB=("vocabulary retention") OR AB=("vocabulary increase") OR AB=("vocabulary improvement") OR AB=("vocabulary development") OR AB=("vocabulary expansion") OR AB=("vocabulary recall") OR AB=("word acquisition") OR AB=("word attainment") OR AB=("word gain") OR AB=("word growth") OR AB=("word knowledge") OR AB=("word learning") OR AB=("word retention") OR AB=("word increase") OR AB=("word improvement") OR AB=("word development") OR AB=("word expansion") OR AB=("word recall") OR AB=("lex* acquisition") OR AB=("lex* attainment") OR AB=("lex* gain") OR AB=("lex* growth") OR AB=("lex* knowledge") OR AB=("lex* learning") OR AB=("lex* retention") OR AB=("lex* increase") OR AB=("lex* improvement") OR AB=("lex* development") OR AB=("lex* expansion") OR AB=("lex* recall") OR AB=("acquisition of vocabulary") OR AB=("attainment of vocabulary") OR AB=("gain in vocabulary") OR AB=("growth in vocabulary") OR AB=("knowledge of vocabulary") OR AB=("learning of vocabulary") OR AB=("retention of vocabulary") OR AB=("increase in vocabulary") OR AB=("improvement in vocabulary") OR AB=("development of vocabulary") OR AB=("expansion of vocabulary") OR AB=("recall of vocabulary") OR AB=("acquisition of word*") OR AB=("attainment of word*") OR AB=("gain in word*") OR AB=("growth in word*") OR AB=("knowledge of word*") OR AB=("learning of word*") OR AB=("retention of word*") OR AB=("increase in word*") OR AB=("improvement in word*") OR AB=("development of word*") OR AB=("expansion of word*") OR AB=("recall of word*") OR AB=("acquisition of lex*") OR AB=("attainment of lex*") OR AB=("gain in lex*") OR AB=("growth in lex*") OR AB=("knowledge of lex*") OR AB=("learning of lex*") OR AB=("retention of lex*") OR AB=("increase in lex*") OR AB=("improvement in lex*") OR AB=("development of lex*") OR AB=("expansion of lex*") OR AB=("recall of lex*"))

Database 4: PsycINFO (via Ovid)

((L1).ab,ti. OR ("first language*").ab,ti. OR ("home language*").ab,ti. OR ("native language*").ab,ti. OR ("mother tongue*").ab,ti. OR ("heritage language*").ab,ti. OR ("community language*").ab,ti. OR ("background language*").ab,ti. OR ("family language*").ab,ti. OR ("ethnic language*").ab,ti. OR ("minority language*").ab,ti. OR (translanguaging).ab,ti. OR (codeswitching).ab,ti. OR ("code-switching").ab,ti. OR ("code switching").ab,ti. OR (codemixing).ab,ti. OR ("code-mixing").ab,ti. OR ("code mixing").ab,ti. OR (codemeshing).ab,ti. OR ("code-meshing").ab,ti. OR ("code meshing").ab,ti.) AND ((L2).ab,ti. OR ("second language*").ab,ti. OR ("additional language*").ab,ti. OR ("non-native language*").ab,ti. OR ("foreign language*").ab,ti. OR ("majority language*").ab,ti. OR ("school language*").ab,ti. OR ("target language*").ab,ti. OR ("academic language*").ab,ti. OR ("mainstream language*").ab,ti. OR ("education* language*").ab,ti. OR ("instructional language*").ab,ti. OR ("language of education").ab,ti. OR ("language of instruction").ab,ti. OR (MFL).ab,ti. OR (ESL).ab,ti. OR (EAL).ab,ti. OR (EFL).ab,ti. OR ("English for speakers of

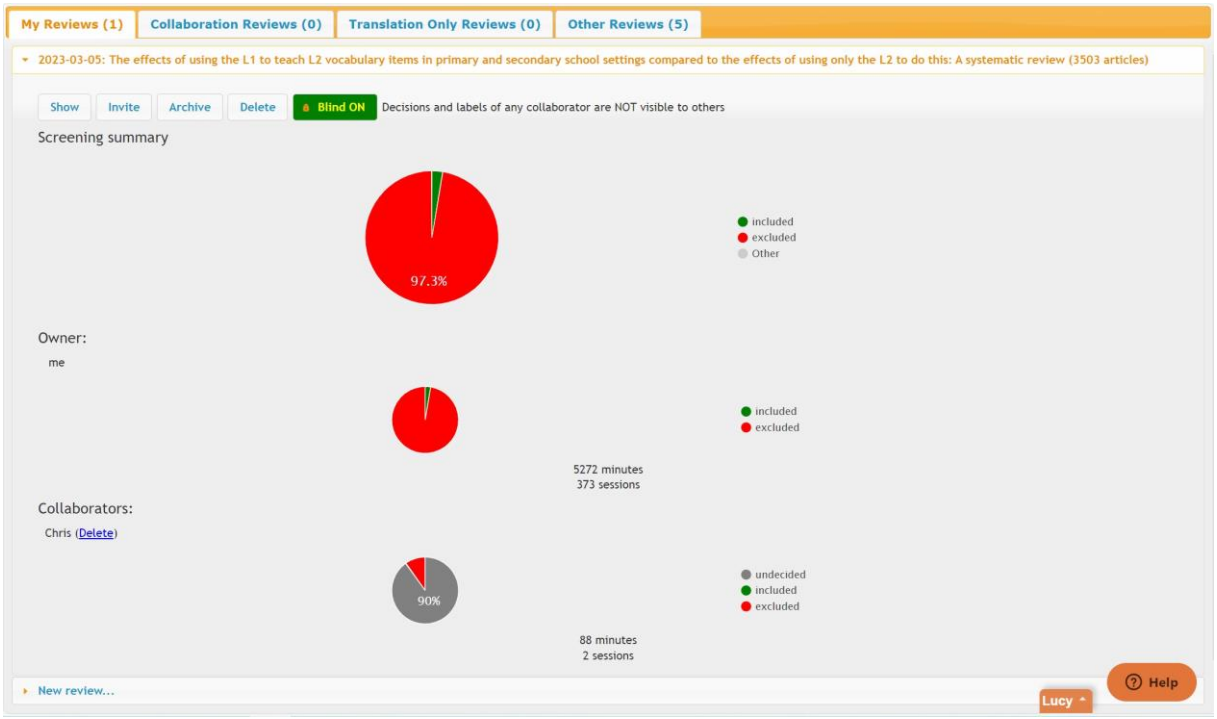
other languages").ab,ti. OR (ESOL).ab,ti. OR ("English language teaching").ab,ti. OR (ELT).ab,ti. OR (TESL).ab,ti. OR (TEAL).ab,ti. OR (TEFL).ab,ti. OR (TESOL).ab,ti. OR ("teaching English to speakers of other languages").ab,ti.) AND (("vocabulary acquisition").ab,ti. OR ("vocabulary attainment").ab,ti. OR ("vocabulary gain").ab,ti. OR ("vocabulary growth").ab,ti. OR ("vocabulary knowledge").ab,ti. OR ("vocabulary learning").ab,ti. OR ("vocabulary retention").ab,ti. OR ("vocabulary increase").ab,ti. OR ("vocabulary improvement").ab,ti. OR ("vocabulary development").ab,ti. OR ("vocabulary expansion").ab,ti. OR ("vocabulary recall").ab,ti. OR ("word acquisition").ab,ti. OR ("word attainment").ab,ti. OR ("word gain").ab,ti. OR ("word growth").ab,ti. OR ("word knowledge").ab,ti. OR ("word learning").ab,ti. OR ("word retention").ab,ti. OR ("word increase").ab,ti. OR ("word improvement").ab,ti. OR ("word development").ab,ti. OR ("word expansion").ab,ti. OR ("word recall").ab,ti. OR ("lex* acquisition").ab,ti. OR ("lex* attainment").ab,ti. OR ("lex* gain").ab,ti. OR ("lex* growth").ab,ti. OR ("lex* knowledge").ab,ti. OR ("lex* learning").ab,ti. OR ("lex* retention").ab,ti. OR ("lex* increase").ab,ti. OR ("lex* improvement").ab,ti. OR ("lex* development").ab,ti. OR ("lex* expansion").ab,ti. OR ("lex* recall").ab,ti. OR ("acquisition of vocabulary").ab,ti. OR ("attainment of vocabulary").ab,ti. OR ("gain in vocabulary").ab,ti. OR ("growth in vocabulary").ab,ti. OR ("knowledge of vocabulary").ab,ti. OR ("learning of vocabulary").ab,ti. OR ("retention of vocabulary").ab,ti. OR ("increase in vocabulary").ab,ti. OR ("improvement in vocabulary").ab,ti. OR ("development of vocabulary").ab,ti. OR ("expansion of vocabulary").ab,ti. OR ("recall of vocabulary").ab,ti. OR ("acquisition of word*").ab,ti. OR ("attainment of word*").ab,ti. OR ("gain in word*").ab,ti. OR ("growth in word*").ab,ti. OR ("knowledge of word*").ab,ti. OR ("learning of word*").ab,ti. OR ("retention of word*").ab,ti. OR ("increase in word*").ab,ti. OR ("improvement in word*").ab,ti. OR ("development of word*").ab,ti. OR ("expansion of word*").ab,ti. OR ("recall of word*").ab,ti. OR ("acquisition of lex*").ab,ti. OR ("attainment of lex*").ab,ti. OR ("gain in lex*").ab,ti. OR ("growth in lex*").ab,ti. OR ("knowledge of lex*").ab,ti. OR ("learning of lex*").ab,ti. OR ("retention of lex*").ab,ti. OR ("increase in lex*").ab,ti. OR ("improvement in lex*").ab,ti. OR ("development of lex*").ab,ti. OR ("expansion of lex*").ab,ti. OR ("recall of lex*").ab,ti.)

Database 5: Scopus

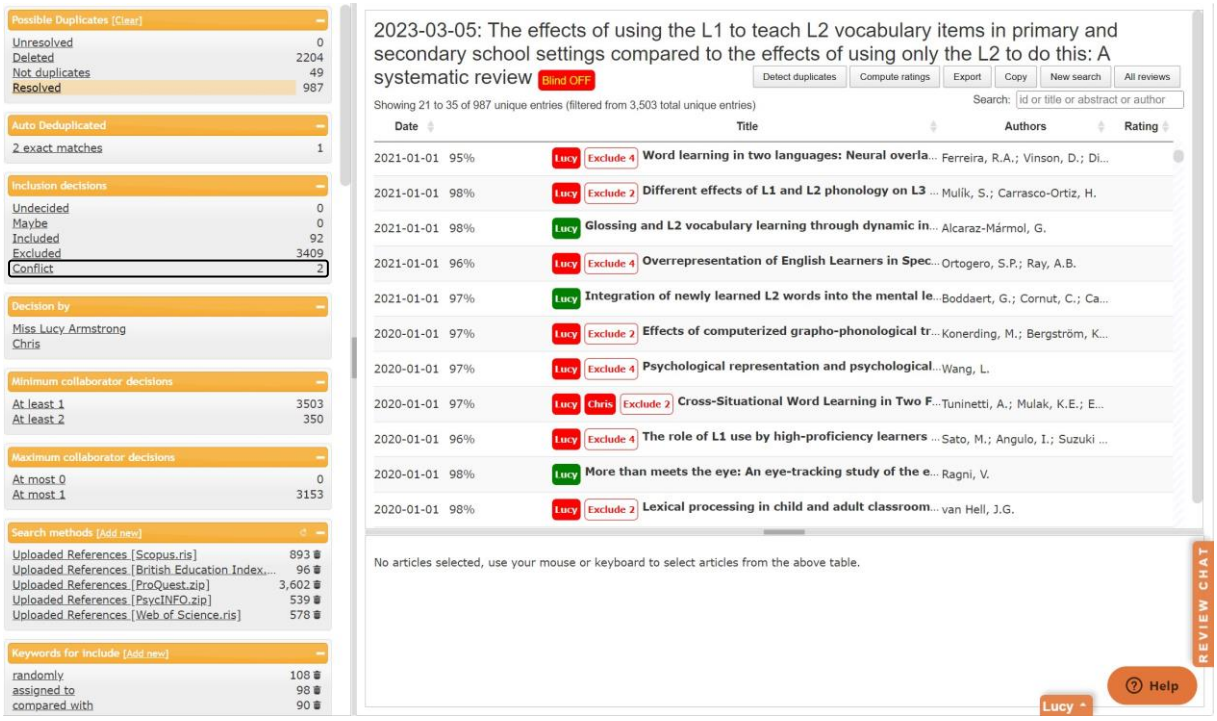
(TITLE-ABS-KEY(L1) OR TITLE-ABS-KEY("first language*") OR TITLE-ABS-KEY("home language*") OR TITLE-ABS-KEY("native language*") OR TITLE-ABS-KEY("mother tongue*") OR TITLE-ABS-KEY("heritage language*") OR TITLE-ABS-KEY("community language*") OR TITLE-ABS-KEY("background language*") OR TITLE-ABS-KEY("family language*") OR TITLE-ABS-KEY("ethnic language*") OR TITLE-ABS-KEY("minority language*") OR TITLE-ABS-KEY(translanguaging) OR TITLE-ABS-KEY(codeswitching) OR TITLE-ABS-KEY("code-switching") OR TITLE-ABS-KEY("code switching") OR TITLE-ABS-KEY(codemixing) OR TITLE-ABS-KEY("code-mixing") OR TITLE-ABS-KEY("code mixing") OR TITLE-ABS-KEY(codemeshing) OR TITLE-ABS-KEY("code-meshing") OR TITLE-ABS-KEY("code meshing")) AND (TITLE-ABS-KEY(L2) OR TITLE-ABS-KEY("second language*") OR TITLE-ABS-KEY("additional language*") OR TITLE-ABS-KEY("non-native language*") OR TITLE-ABS-KEY("foreign language*") OR TITLE-ABS-KEY("majority language*") OR TITLE-ABS-KEY("school language*") OR TITLE-ABS-KEY("target language*") OR TITLE-ABS-KEY("academic language*") OR TITLE-ABS-KEY("mainstream language*") OR TITLE-ABS-KEY("education* language*") OR TITLE-ABS-KEY("instructional language*") OR TITLE-ABS-KEY("language of education") OR TITLE-ABS-KEY("language of instruction") OR TITLE-ABS-KEY(MFL) OR TITLE-ABS-KEY(ESL) OR TITLE-ABS-KEY(EAL) OR

TITLE-ABS-KEY(EFL) OR TITLE-ABS-KEY("English for speakers of other languages") OR TITLE-ABS-KEY(ESOL) OR TITLE-ABS-KEY("English language teaching") OR TITLE-ABS-KEY(ELT) OR TITLE-ABS-KEY(TESL) OR TITLE-ABS-KEY(TEAL) OR TITLE-ABS-KEY(TEFL) OR TITLE-ABS-KEY(TESOL) OR TITLE-ABS-KEY("teaching English to speakers of other languages")) AND (TITLE-ABS-KEY("vocabulary acquisition") OR TITLE-ABS-KEY("vocabulary attainment") OR TITLE-ABS-KEY("vocabulary gain") OR TITLE-ABS-KEY("vocabulary growth") OR TITLE-ABS-KEY("vocabulary knowledge") OR TITLE-ABS-KEY("vocabulary learning") OR TITLE-ABS-KEY("vocabulary retention") OR TITLE-ABS-KEY("vocabulary increase") OR TITLE-ABS-KEY("vocabulary improvement") OR TITLE-ABS-KEY("vocabulary development") OR TITLE-ABS-KEY("vocabulary expansion") OR TITLE-ABS-KEY("vocabulary recall") OR TITLE-ABS-KEY("word acquisition") OR TITLE-ABS-KEY("word attainment") OR TITLE-ABS-KEY("word gain") OR TITLE-ABS-KEY("word growth") OR TITLE-ABS-KEY("word knowledge") OR TITLE-ABS-KEY("word learning") OR TITLE-ABS-KEY("word retention") OR TITLE-ABS-KEY("word increase") OR TITLE-ABS-KEY("word improvement") OR TITLE-ABS-KEY("word development") OR TITLE-ABS-KEY("word expansion") OR TITLE-ABS-KEY("word recall") OR TITLE-ABS-KEY("lex* acquisition") OR TITLE-ABS-KEY("lex* attainment") OR TITLE-ABS-KEY("lex* gain") OR TITLE-ABS-KEY("lex* growth") OR TITLE-ABS-KEY("lex* knowledge") OR TITLE-ABS-KEY("lex* learning") OR TITLE-ABS-KEY("lex* retention") OR TITLE-ABS-KEY("lex* increase") OR TITLE-ABS-KEY("lex* improvement") OR TITLE-ABS-KEY("lex* development") OR TITLE-ABS-KEY("lex* expansion") OR TITLE-ABS-KEY("lex* recall") OR TITLE-ABS-KEY("acquisition of vocabulary") OR TITLE-ABS-KEY("attainment of vocabulary") OR TITLE-ABS-KEY("gain in vocabulary") OR TITLE-ABS-KEY("growth in vocabulary") OR TITLE-ABS-KEY("knowledge of vocabulary") OR TITLE-ABS-KEY("learning of vocabulary") OR TITLE-ABS-KEY("retention of vocabulary") OR TITLE-ABS-KEY("increase in vocabulary") OR TITLE-ABS-KEY("improvement in vocabulary") OR TITLE-ABS-KEY("development of vocabulary") OR TITLE-ABS-KEY("expansion of vocabulary") OR TITLE-ABS-KEY("recall of vocabulary") OR TITLE-ABS-KEY("acquisition of word*") OR TITLE-ABS-KEY("attainment of word*") OR TITLE-ABS-KEY("gain in word*") OR TITLE-ABS-KEY("growth in word*") OR TITLE-ABS-KEY("knowledge of word*") OR TITLE-ABS-KEY("learning of word*") OR TITLE-ABS-KEY("retention of word*") OR TITLE-ABS-KEY("increase in word*") OR TITLE-ABS-KEY("improvement in word*") OR TITLE-ABS-KEY("development of word*") OR TITLE-ABS-KEY("expansion of word*") OR TITLE-ABS-KEY("recall of word*") OR TITLE-ABS-KEY("acquisition of lex*") OR TITLE-ABS-KEY("attainment of lex*") OR TITLE-ABS-KEY("gain in lex*") OR TITLE-ABS-KEY("growth in lex*") OR TITLE-ABS-KEY("knowledge of lex*") OR TITLE-ABS-KEY("learning of lex*") OR TITLE-ABS-KEY("retention of lex*") OR TITLE-ABS-KEY("increase in lex*") OR TITLE-ABS-KEY("improvement in lex*") OR TITLE-ABS-KEY("development of lex*") OR TITLE-ABS-KEY("expansion of lex*") OR TITLE-ABS-KEY("recall of lex*"))

Appendix 5a: A Screenshot of Rayyan Showing Double Screening



Appendix 5b: A Screenshot of Rayyan Showing Two Conflicting Decisions



Appendix 6: The Eligibility Outcomes of Full-Text Screening

Study Reference	Eligibility Outcome
Ajabshir (2022)	Excluded
Al Hashim and Carstens (2015)	Excluded
Al Ramadhan (2020)	Included
Alcaraz-Mármol (2021)	Included
Alghamdi (2016)	Excluded
Alghamdi (2018)	Excluded
Alharthi (2018)	Excluded
Alsadoon (2021)	Excluded
Amirian et al. (2016)	Included
Arpacı (2016)	Included
Arvizu (2020)	Included
Asiyaban and Bagheri (2012)	Excluded
Bairstow and Lavaur (2017)	Excluded
Baltova (1999)	Included
Bisson et al. (2015)	Excluded
Boddaert et al. (2021)	Excluded
Busse et al. (2021)	Included
Camó and Ballester (2015)	Included
Carpenter and Olson (2012)	Excluded
Cekic (2022)	Excluded
Chen (2013)	Excluded
Chen et al. (2020)	Excluded
Choi et al. (2016)	Included
Christensen et al. (2007)	Excluded
Curuana et al. (2014)	Excluded
Dabaghi and Rafiee (2012)	Excluded
del Pilar Agustín Llach (2009)	Included
Dizon and Thanyawatpokin (2021)	Excluded
Dolgunsöz and Kimsesiz (2021)	Excluded
Ferguson (2009)	Excluded
Galante (2020)	Excluded
Gallardo (2021)	Excluded
Gallardo (2023)	Excluded
García-Gámez and Macizo (2022)	Excluded
Gettys et al. (2001)	Excluded
Ghabanchi and Ayoubi (2012)	Included
Ghahari and Heidarolad (2015)	Excluded
Grace (1998)	Excluded
Grace (2000)	Excluded
Gyllstad et al. (2023)	Included
Hashemzadeh (2012)	Excluded
He (2019)	Included

Hopp et al. (2021)	**Excluded**
Hsieh (2020)	**Excluded**
Hu et al. (2014)	**Included**
Hulstijn et al. (1996)	**Excluded**
Ina (2014)	**Included**
Jin and Webb (2020)	**Excluded**
Kaushik (2013)	**Excluded**
Kongtawee and Sappapan (2018)	**Included**
Koolstra and Beentjes (1999)	**Included**
Laufer and Girsai (2008)	**Included**
Laufer and Shmueli (1997)	**Included**
Lee and Macaro (2013)	**Included**
Lugo-Neris et al. (2010)	**Excluded**
Mazur et al. (2016)	**Excluded**
Muñoz et. al. (2023)	**Included**
Ollerhead et al. (2020)	**Excluded**
Ong and Zhang (2018)	**Excluded**
Peters (2019)	**Included**
Peters et al. (2016)	**Included**
Pressley et al. (1981)	**Excluded**
Prince (1996)	**Excluded**
Pujadas and Muñoz (2019)	**Included**
Ragni (2020)	**Excluded**
Ramachandran and Rahim (2004)	**Included**
Rassaei (2020)	**Excluded**
Read et al. (2021)	**Excluded**
Rott (2007)	**Excluded**
Rouhi and Mohebbi (2012)	**Excluded**
Sadeghi and Khezrlou (2012)	**Excluded**
Samian et al. (2016)	**Excluded**
Sawada (2009)	**Included**
Saz et al. (2015)	**Excluded**
Shabani and Rahimy (2020)	**Excluded**
Soleimani and Esmaeili (2012)	**Excluded**
Tadayonifar et al. (2021)	**Excluded**
Teng (2020)	**Excluded**
Tonzar et al. (2009)	**Included**
Torang and Weisi (2023)	**Excluded**
Vela (2015)	**Excluded**
Wang and Pellicer-Sánchez (2022)	**Excluded**
Williams et al. (2019)	**Excluded**
Wu and Yang (2022)	**Included**
Yoshii (2006)	**Excluded**
Zhang (2022)	**Included**
Zhang and Graham (2020)	**Included**

Appendix 7: An Example Data Extraction Form

Data Extraction Form for Study 37

Note: Throughout this form, only information relevant to the review will be recorded. In other words, should a study report on multiple experiments, then only those involving typical subjects from mainstream primary or secondary educational environments, and investigating the effects of L1-mediated vocabulary teaching approaches in comparison to L2-mediated vocabulary teaching approaches on L2 vocabulary acquisition, will be acknowledged. In line with this, any outcomes not pertaining to L2 vocabulary acquisition will not be recorded, and nor will the research questions, data or findings that relate to these alternative outcomes.

Section A: Administrative Information

Full Reference: Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008-1032. https://doi.org/10.1002/tesq.531.

In-text Reference	Peters (2019)
Study Title	The Effect of Imagery and On-Screen Text on Foreign Language Vocabulary Learning From Audiovisual Input
Identification Method *(Database searching, backwards and forwards citation, or already known to the reviewer, etc.)*	Database searching
Publication Type	Journal article
Publication Language	English

Section B: Contextual Information

Study Location	Flanders, Belgium
Educational Setting	Grades 5 and 6 of secondary school
Type of Language Programme	English as a Foreign Language
Study Participants *(Their number, ages, genders, first languages and any attrition)*	142 Dutch L1 speakers, aged 16.4 (on average), but as 23 of them did not provide a complete dataset, and 1 of them had a much lower score on the post-test than the pre-test, the sample was reduced to 118 Dutch L1 speakers *(Note 1: The participants came from four grade 5 classes and two grade 6 classes at a secondary school in the Dutch-speaking part of Belgium)* *(Note 2: The gender of the participants was not provided)* Additional Information • Socio-economic status: Not stated • Time spent learning English: Learners in the fifth grade were in their fourth year of formal English language instruction, and learners in the sixth grade were in their fifth year of formal English language instruction

	• Current proficiency level: Intermediate (B1 on the CEFR) *(Note: This level is recorded using the terminology of the study)*

Section C: Intervention Information

Study Design *(Relevant to this review)*	A pre-test/post-test design
Research Question(s)	Research Question 2: Does input modality (L1 subtitles, L2 captions, no subtitles) have an effect on word learning from audiovisual input? Research Question 3: Is there an interaction effect of imagery and input modality on word learning from audiovisual input?
Number of Conditions *(Both experimental and control)*	Three conditions
Names and Sizes of Each Condition *(Including significant information about the participants)*	1. The L1 subtitles condition (41 participants) 2. The L2 captions condition (36 participants) 3. The no subtitles (control) condition (41 participants) *(Note: According to the results of the Vocabulary Size Test, and the ANVOA and Bonferroni post-hoc tests that were conducted on them, participants in the L2 captions condition differed significantly from those in the no subtitles condition (p = .008) in terms of their general vocabulary knowledge)* **TABLE 3** **Descriptive Statistics for the Vocabulary Size Test, per Group** *N* / Mean (*SD*) (Max = 140) / [95% confidence interval] Captions group: 36 / 94.89 (10.53) / [91.33, 98.45] L1 subtitles group: 41 / 91.95 (9.74) / [88.88, 95.02] No subtitles group: 41 / 88.27 (8.15) / [85.7, 90.84]
Method of Participant Allocation to Each Condition	Students from six classes – four in the fifth grade and two in the sixth grade – were randomly assigned to one of the two experimental conditions, or to the control condition. *(Note: Whether these were intact classes, and how the random assignment of participants was conducted, is not confirmed)*
L1-mediated Vocabulary Teaching Approach(es)	Participants in the L1 subtitles condition were shown the 12-minute *Planet Earth* documentary excerpt twice, with L2 English audio and L1 Dutch subtitles on both occasions. *(Note: They were told that they would have to respond to comprehension questions after the viewing, but not that they would be required to complete a vocabulary post-test)*
L2-mediated Vocabulary Teaching Approach(es)	Participants in the L2 captions condition were also shown the 12-minute *Planet Earth* documentary excerpt twice, with L2 English audio and L2 English captions on both occasions. Participants in the no subtitles condition were shown the 12-minute *Planet Earth* documentary excerpt twice, but with L2 English audio only, and no subtitles or captions.
Administrator of Each Condition	Not stated.
Target L2 Vocabulary	36 words embedded in the script of an excerpt (11 minutes and 25 seconds long) from the English language documentary

	Planet Earth, which was chosen for its high levels of imagery. *(Note 1: According to the VocabProfile tool in Lextutor, 90% of the lexis in the excerpt comes from the 3,000 most frequent word families in English, and 95% from the 5,000 most frequent word families)* *(Note 2: 15 of the 36 words had imagery support, while 21 did not)* *(Note 3: 10 of the 36 words were cognates with Dutch, while 26 were not)* *(Note 4: Frequency of word occurrence in the script ranged from 1 to 9, and some of the words were high-frequency in general, included for reasons of test motivation, while others were low frequency in general)* *(Note 5: The concreteness of the words ranged from 1.71 to 4.97, according to Brysbaert, Warriner, and Kuperman's (2014) concreteness ratings)*
Study Duration	Two 50-minute sessions held one week apart. *(Note: In the first of these sessions, participants completed the Vocabulary Size Test and the pre-test, and in the second, they watched the documentary excerpt and took the post-test)*

Section D: Outcome Information

Outcome Measure(s) *(Relevant to this review)*	Receptive knowledge of the target L2 words. Post-test (immediately after the treatment): The two-part pre-test was re-administered to the participants. *(Note: This post-test had an acceptable reliability, with a Cronbach's alpha score of .74 for the form recognition part, and .74 for the meaning recall part)*
Pre-Test Measure(s)	Receptive knowledge of the target L2 words. Pre-test (one week before the treatment): In this two-part pre-test, which assessed both form recognition and meaning recall, the participants were first exposed to the spoken and written form of 44 L2 words – 36 of which were the target words, and 8 of which were non-word distractors – and asked to indicate whether they had seen or heard them before by circling 'yes' or 'no'. Following this, they were then asked to give all known meanings of the 44 L2 words, be they in the form of L1 translations, L2 synonyms or L2 definitions. *(Note 1: If participants indicated incorrectly that they recognised three or more non-words, then their data was not included in analysis of the form recognition part of the pre-test, and this was the case for 23 of them)* *(Note 2: This pre-test had an acceptable reliability, with a Cronbach's alpha score of .70 for the form recognition part, and .72 for the meaning recall part)* Participants' general knowledge of L2 vocabulary. The Vocabulary Size Test (one week before the treatment): Participants were asked to complete the entirety of the VST by Nation and Beglar (2007), which presents 10 words from each of the first 14 frequency bands of 1,000 word families in English in multiple-choice format, and asks them to identify their correct definition from four possible options. *(Note: The purpose of this test was to ascertain the overall vocabulary knowledge level of the participants in each condition, and its reliability was acceptable, with a Cronbach's alpha of .80)*

Quantitative Data for Each Measure In Each Condition	Descriptive and inferential statistics: **TABLE 4** **Descriptive Statistics Form and Meaning Pretest/Posttest, per Group** 		Form pretest mean (*SD*) [95% CI]	Form posttest mean (*SD*) [95% CI]	Meaning pretest mean (*SD*) [95% CI]	Meaning posttest mean (*SD*) [95% CI]
---	---	---	---	---
Captions	26.19 (3.70) [24.70, 27.69]	29.00 (3.12) [27.74, 30.26]	17.08 (4.35) [15.61, 18.56]	21.06 (4.29) [19.60, 22.51]
L1 subtitles	25.53 (3.88) [24.13, 26.93]	26.38 (3.93) [24.96, 27.79]	15.76 (3.57) [14.66, 16.85]	18.61 (3.62) [17.47, 19.75]
No subtitles	24.11 (3.84) [22.83, 25.39]	24.51 (3.06) [23.49, 25.53]	14.24 (2.70) [13.39, 15.10]	16.93 (3.33) [15.88, 17.98]	 *Note.* The Form test is based on data of 95 participants, the Meaning test on data of 118 participants. CI = confidence interval; Max score possible = 36. **TABLE 5** **Parameter Estimates of Generalized Estimating Equations for Form Recognition Test** 	Parameter	*B*	SE	95% Wald CI		Wald chi-square	*p*	Exp (B)	95% Wald CI for Exp(B)	
---	---	---	---	---	---	---	---	---	---
(Intercept)	−2.55	.92	−4.36	−.75	7.71	.006	0.08	0.01	0.47
Group = Captions	.90	.27	.38	1.42	11.35	.001	2.45	1.46	4.13
Group = L1Subtitles	−.03	.27	−.55	.49	.01	.914	0.97	0.58	1.64
Group = NoSubtitles	0								
Imagery = 0	−1.04	.25	−1.52	−.56	17.88	.000	0.36	0.22	0.57
Imagery = 1	0								
Cognate = 0	−1.91	.23	−2.36	−1.46	69.43	.000	0.15	0.09	0.23
Cognate = 1	0								
VST	.02	.01	.01	.04	6.06	.014	1.03	1.01	1.05
Corpus frequency	.51	.13	.25	.77	14.56	.000	1.66	1.28	2.15
FreqOccurrence	.23	.05	.12	.33	18.27	.000	1.25	1.13	1.39
Group = Captions X Imagery = 0	.56	.34	−.11	1.24	2.67	.102	1.75	0.89	3.44
Group = L1Subtitles X Imagery = 0	.61	.34	−.06	1.29	3.16	.076	1.84	.94	3.62 | *Note.* CI = confidence interval; FreqOccurrence = frequency of occurrence; VST = score on Vocabulary Size Test. **TABLE 6** **Parameter Estimates of Generalized Estimating Equations for Meaning Recall Test** | Parameter	*B*	SE	95% Wald CI		Wald chi-square	*p*	Exp (B)	95% Wald CI for Exp(B)	
---	---	---	---	---	---	---	---	---	---
(Intercept)	−4.40	.63	−5.63	−3.17	49.07	.000	0.01	0	0.04
Group = Captions	0.68	.15	0.39	0.96	21.73	.000	1.97	1.48	2.62
Group = L1Subtitles	0.12	.16	−0.20	0.44	0.56	.460	1.13	0.82	1.55
Group = NoSubtitles	0								
Imagery = 0	−1.10	.14	−1.37	−0.84	65.64	.000	0.33	0.25	0.43
Imagery = 1	0								
Cognate = 0	−1.95	.14	−2.23	−1.66	181.61	.000	0.14	0.11	0.19
Cognate = 1	0								
VST	0.04	.01	0.02	0.05	27.52	.000	1.04	1.02	1.05
Corpus frequency	0.66	.10	0.46	0.86	41.60	.000	1.94	1.59	2.37
FreqOccurrence	1.17	.03	0.06	0.17	17.23	.000	1.12	1.06	1.19	 *Note.* CI = confidence interval; FreqOccurrence = frequency of occurrence; VST = score on Vocabulary Size Test.
Effect Size(s)	Not stated
Study Finding(s)	The findings of this study showed that participants in the L2 captions condition made the most vocabulary learning gains, in terms of both form recognition and meaning recall. These gains were slightly higher for the former, but still statistically significant for both, and on-screen imagery was also revealed to be a facilitative factor in L2 vocabulary learning, although there was no significant interaction between experimental group and imagery, which means that the effect of imagery was not dependent on the type of subtitles provided.

Note: Copies of the data extraction forms for the other 44 studies are available upon request.

Appendix 8: The Adapted EPHPP Risk of Bias Assessment Tool

Risk of Bias Assessment for Study 37

Full Reference: Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008-1032. https://doi.org/10.1002/tesq.531.

In-text reference: Peters (2019).

Component Ratings

A) Selection Bias

(Q1) Are the individuals selected to participate in the study likely to be representative of the target population?

1 Very likely
2 Somewhat likely
3 Not likely
4 Can't tell

(Q2) What percentage of selected individuals agreed to participate?

1 80-100% agreement
2 60-79% agreement
3 less than 60% agreement
4 Not applicable
5 Can't tell

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

B) Study Design

(Q1) Indicate the study design:

1 Randomised controlled trial (with cluster randomization)
2 Non-randomised comparison (comparison groups prospectively generated by a means other than random allocation)
3 Cohort analytic (two group pre + post)
4 Case-control
5 Cohort (one group pre + post (before and after))
6 Interrupted time series
7 Other, specify: _____
8 Can't tell

(Q2) Was the study described as randomised? If NO, go to Component C.

No Yes

If YES, was the method of randomisation described? (See dictionary)

No Yes

If YES, was the method appropriate? (See dictionary)

No Yes

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

C) Confounders

(Q1) Were there important differences between groups prior to the intervention that were known to the researcher(s)?

1 Yes
2 No
3 Can't tell

The following are examples of confounders: (1) Race, (2) Sex, (3) Marital status/family, (4) Age, (5) Socio-economic status (income or class), (6) Education, (7) Health status, (8) Pre-intervention score on outcome measure

(Q2) If YES, indicate the percentage of relevant confounders that were controlled (either in the design (e.g. stratification, matching) or analysis)?

1 80-100% (most)
2 60-79% (some)
3 Less than 60% (few or none)
4 Can't Tell

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

D) Blinding

(Q1) Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?

1 Yes
2 No
3 Can't tell

(Q2) Were the study participants aware of the research question(s)?

1 Yes
2 No
3 Can't tell

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

E) Data Collection Methods

(Q1) Were data collection tools shown to be valid?

1 Yes
2 No
3 Can't tell

(Q2) Were data collection tools shown to be reliable?

1 Yes
2 No
3 Can't tell

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

F) Withdrawals and Drop-outs

(Q1) Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?

1 Yes
2 No
3 Can't tell
4 Not applicable (i.e. one-time surveys or interviews)

(Q2) Indicate the percentage of participants completing the study. (If the percentage differs by groups, record the lowest).

1 80%-100%
2 60%-79%
3 Less than 60%
4 Can't tell
5 Not applicable (i.e. retrospective case-control)

Rate This Section	STRONG	MODERATE	WEAK
(See dictionary)	1	2	3

Global Rating

Component Ratings

Please transcribe the information from the gray boxes on pages 1-3 onto this page. Then see the dictionary on how to rate this section.

A	**Selection Bias**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	
B	**Study Design**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	
C	**Confounders**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	
D	**Blinding**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	
E	**Data Collection Methods**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	
F	**Withdrawals and Drop-outs**	**STRONG**	**MODERATE**	**WEAK**	
		1	2	3	**NOT APPLICABLE**

GLOBAL RATING FOR THIS PAPER (circle one):

1	STRONG	(no WEAK ratings)
2	MODERATE	(one WEAK rating)
3	WEAK	(two or more WEAK ratings)

If more than one reviewer assigns the ratings, complete the following section:

Is there a discrepancy between the two reviewers with respect to the component (A-F) ratings?

No Yes

If YES, indicate the reason for the discrepancy:

1 Oversight
2 Differences in interpretation of criteria
3 Differences in interpretation of study

Final decision of both reviewers (circle one): 1 STRONG
2 MODERATE
3 WEAK

(Adapted from the Quality Assessment Tool for Quantitative Studies by the EPHPP)

Appendix 9: The Reference List of Included Reports

S1: Al Nassir, M. E. (2012). *Meaning recall and retention: Comparison between translation method and pictorial method in learning vocabulary in Saudis' school* (Publication No. 1510943) [Master's thesis, Colorado State University]. ProQuest Dissertations Publishing.

S2: Al Ramadhan, M. H. (2020). L1 textual glosses and word repetition: Facilitative interventions for incidental vocabulary acquisition. *International Journal of Instruction, 13*(4), 815-832. https://doi.org/10.29333/iji.2020.13450a.

S3: Alcaraz-Mármol, G. (2021). Glossing and L2 vocabulary learning through dynamic instruction in the context of primary education. *Revista de Lingüística y Lenguas Aplicadas, 16*(1), 1-10. http://doi.org/10.4995/rlyla.2021.14691.

S4: Alsuhaymi, O. H. M. (2019). Impact of using pictures on English vocabulary retention and recall by primary stage learners in Al-Mikhwah Province. *Journal of Educational and Psychological Sciences, 3*(16), 140-153. https://doi.org/10.26389/AJSRP.O311218.

S5: Amirian, S. M. R., Zareian, G., & Nour, S. (2016). Comparing the effects of L1 and L2 definition on incidental vocabulary learning through listening to stories. *Journal of Language Teaching and Research, 7*(3), 604-612. https://doi.org/10.17507/jltr.0703.24.

S6: Arpacı, D. (2016). The effects of accessing L1 versus L2 definitional glosses on L2 learners' reading comprehension and vocabulary learning. *Eurasian Journal of Applied Linguistics, 2*(1), 15-29. https://doi.org/10.32601/ejal.460988.

S7: Arvizu, M. N. G. (2020). L2 vocabulary acquisition through narratives in an EFL public elementary school. *IAFOR Journal of Education: Language Learning in Education, 8*(1), 115-128. https://doi.org/10.22492/ije.8.1.07.

S8: Baltova, I. (1999). *The effect of subtitled and staged video input on the learning and retention of content and vocabulary in a second language* (Publication No. 41096) [Doctoral thesis, University of Toronto]. ProQuest Dissertations Publishing.

S9: Baranowska, K. (2020). Learning most with least effort: Subtitles and cognitive load. *ELT Journal, 74*(2), 105-115. https://doi.org/10.1093/elt/ccz060.

S10: Bueno-Alastuey, M. C., & García, D. N. (2019). The effectiveness of translation, image-based and video-based methodologies for receptive and productive vocabulary acquisition. https://doaj.org/article/bed43ead504b4e638d097d45a3370575.

S11: Busse, V., Cenoz, J., Dalmann, N., & Rogge, F. (2020). Addressing linguistic diversity in the language classroom in a resource-oriented way: An intervention study with primary school children. *Language Learning, 70*(2), 382-419. https://doi.org/10.1111/lang.12382.

S12: Busse, V., McLaren, L., & Dahm, A. (2021). Responding to migration-related diversity in the classroom: A comparison of diversity-sensitive approaches to stimulate word acquisition

in early FL teaching. *Journal of Multilingual and Multicultural Development,* 1-20. https://doi.org/10.1080/01434632.2021.2005611.

S13: Camó, A. C., & Ballester, E. P. (2015). The effects of using L1 translation on young learners' foreign language vocabulary learning. *Estudios de Lingüística Inglesa Aplicada, 15*(1), 109-134. https://doi.org/10.12795/elia.2015.i15.06.

S14: Chalmers, H. (2019). *Leveraging the L1: The role of EAL learners' first language in their acquisition of English vocabulary* [Unpublished doctoral thesis, Oxford Brookes University]. https://doi.org/10.13140/RG.2.2.20764.13445

S15: Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences, 45*(1), 137-143. http://dx.doi.org/10.1016/j.lindif.2015.11.018.

S16: Del Pilar Agustín Llach, M. (2009). The effect of reading, reading and comprehension, and sentence writing in lexical learning in a foreign language: Some preliminary results. *RESLA, 22*(1), 9-33. https://dialnet.unirioja.es/servlet/articulo?codigo=3138255.

S17: Emirmustafaoğlu, A., & Gökmen, D. U. (2015). The effects of picture vs. translation mediated instruction on L2 vocabulary learning. *Procedia – Social and Behavioral Sciences 199*(1), 357-362. https://doi.org/10.1016/j.sbspro.2015.07.559.

S18: Ghabanchi, Z., & Ayoubi, E. S. (2012). Incidental vocabulary learning and recall by intermediate foreign language students: The influence of marginal glosses, dictionary use, and summary writing. *Journal of International Education Research, 8*(2), 85-96. https://doi.org/10.19030/jier.v8i2.6828.

S19: Gyllstad, H., Sundqvist, P., Sandlund, E., & Källkvist, M. (2022). Effects of word definitions on meaning recall: A multisite intervention in language-diverse second language English classrooms. *Language Learning, 73*(2), 403-444. https://doi.org/10.1111/lang.12527.

S20: He, L. (2019). How do different keyword captioning strategies impact students' performance in oral and written production tasks? A pilot study. *The EUROCALL Review, 27*(1), 14-23. https://doi.org/10.4995/eurocall.2019.11371.

S21: Helle, M. (2014). *The influence of the subtitled mode on the acquisition of English vocabulary* [Unpublished Master's thesis, University of Bergen].

S22: 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. https://doi.org/10.1080/09571736.2013.849751.

S23: Hsu, C.-K., Hwang, G.-J., Chang, Y.-T., & Chang, C.-K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society, 16*(1), 403-414.

S24: Hu, S., Vongpumivitch, V., Chang, J. S., & Liou, H. (2014). The effects of L1 and L2 e-glosses on incidental vocabulary learning of junior high-school English students. *ReCALL, 26*(1), 80-99. https://doi.org/10.1017/S0958344013000244.

S25: Ina, L. (2014). Incidental foreign-language acquisition by children watching subtitled television programs. *The Turkish Online Journal of Educational Technology, 13*(4), 81-87.

S26: Kongtawee, P., & Sappapan, P. (2018). The effects of L1 and L2 hypertext glosses on reading comprehension and vocabulary retention among Thai secondary school students. *Arab World of English Journal, 9*(3), 367-380. https://dx.doi.org/10.24093/awej/vol9no3.24.

S27: Koolstra, C. M., & Beentjes, J. W. J. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development, 47*(1), 51-60. https://doi-org.ezproxy-prd.bodleian.ox.ac.uk/10.1007/BF02299476.

S28: Laufer, B., & Girsai, N. (2008a). The use of native language for improving second language vocabulary: an exploratory study. In A. Stavans & I. Kupferberg (Eds.), *Studies in language and language education (essays in honor of Elite Olshtain)* (pp. 261-275). The Hebrew University Magnes Press.

S29: Laufer, B., & Girsai, N. (2008b). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716. https://doi.org/10.1093/applin/amn018.

S30: Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*(1), 89-108. https://doi.org/10.1177/003368829702800106.

S31: Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal, 97*(4), 887-901. https://doi.org/10.1111/j.1540-4781.2013.12044.x.

S32: López, A. (2016). *Comparison between the effectiveness of code-mixing and providing L2 definitions in EFL vocabulary acquisition* [Unpublished Master's dissertation, Universidad Pública de Navarra]. Academica-e. https://hdl.handle.net/2454/21499.

S33: Lwo, L., & Lin, M. C. (2012). The effects of captions in teenagers' multimedia L2 learning. *ReCALL, 24*(2), 188-208. https://doi.org/10.1017/S0958344012000067.

S34: Miyasako, N. (2002). Does text-glossing have any effects on incidental vocabulary learning through reading for Japanese senior high school students? *Language Education and Technology, 39*(1), 1-20. https://doi.org/10.24539/let.39.0_1.

S35: Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research, 39*(1), 13-37. https://doi.org/10.1177/02676583211015797.

S36: Peters, E., Heynen, E., & Puimege, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System, 63*(1), 134-148. http://dx.doi.org/10.1016/j.system.2016.10.002.

S37: Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008-1032. https://doi.org/10.1002/tesq.531.

S38: 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. https://doi.org/10.1080/09571736.2019.1616806.

S39: Ramachandran, S. D., & Rahim, H. A. (2004). Meaning recall and retention: The impact of the translation method on elementary level learners' vocabulary learning. *RELC Journal, 35*(2), 161-178. https://doi.org/10.1177/003368820403500205.

S40: Sawada, K. (2009). *Vocabulary acquisition through listening and its relation to learning channel preferences* (Publication No. 3371998) [Doctoral thesis, Temple University]. ProQuest Dissertations Publishing.

S41: Sieh, Y. (2008). A possible role for the first language in young learners' processing and storage of foreign language vocabulary. *Annual Review of Education, Communication, and Language Sciences, 5*(1), 136-160.

S42: Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning, 59*(3), 623-646. https://doi.org/10.1111/j.1467-9922.2009.00519.x.

S43: Wu, H., & Yang, X. (2022). Effectiveness of textually-enhanced captions on Chinese high-school EFL learners' incidental vocabulary. *Porta Linguarum, 38*(1), 209-228. https://doi.org/10.30827/portalin.vi38.23511.

S44: Zhang, P., & Graham, S. (2020). Vocabulary learning through listening: Comparing L2 explanations, teacher codeswitching, contrastive focus-on-form and incidental learning. *Language Teaching Research, 24*(6), 765-784. https://doi.org/10.1177/1362168819829022.

S45: Zhang, P. (2022). How does repetition affect vocabulary learning through listening to the teacher's explicit instruction? The moderating role of listening proficiency and preexisting vocabulary knowledge. *Language Teaching Research*, 1-24. https://doi.org/10.1177/13621688221140521.

Appendix 10: A Summary of Key Descriptive Data from the Included Studies

Study ID	Study Reference	Study Location	Study Context	Study Design	Study Subjects	Study Focus	Outcome Measures	Study Duration	Summary Finding
S1	Al Nassir (2012)	Saihat, Saudi Arabia	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	36 Arabic L1 speakers, aged 17 (on average)	The L1 translation method vs. the pictorial method	Writing: Productive knowledge of 45 target L2 words	Around a week and a half	Did not favour L1 use
S2	Al Ramadhan (2020)	Saudi Arabia	Secondary school EFL programme	Non-randomised comparison	146 Arabic L1 speakers, aged 15-17	Reading with L1 glossing vs. L2 word repetition vs. L1 glossing and L2 word repetition vs. no glossing or repetition	Speaking and writing: Productive and receptive knowledge of 10 target L2 words	Not stated	Favoured L1 use
S3	Alcaraz-Mármol (2021)	The Latina district of Madrid, Spain	Primary school EFL programme	Non-randomised comparison	60 Spanish L1 speakers, aged 11-12	Reading with L1 glossing vs. L1 or L2 definition discussion vs. no definitions	Writing: Productive and receptive knowledge of 10 target L2 words	Around three weeks	Did not favour L1 use
S4	Alsuhaymi (2019)	Al-Mikhwah province, Saudi Arabia	Primary school EFL programme	Randomised controlled trial (with cluster randomisation)	36 Arabic L1 speakers, aged 11 (on average)	The L1 translation method vs. the pictorial method	Reading: Receptive knowledge of 30 target L2 words	At least two weeks	Did not favour L1 use

S5	Amirian et al. (2016)	Sabzevar, Iran	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	74 Persian L1 speakers, aged 17-18	Listening with L1 glossing vs. L2 glossing vs. L1 and L2 glossing vs. no glossing	Reading and writing: Productive and receptive knowledge of 92 target L2 words	Just over six weeks	No difference detected
S6	Arpacı (2016)	Ankara, Turkey	Secondary school EFL programme	Randomised controlled trial (with individual randomisation)	81 Turkish L1 speakers	Reading with L1 glossing vs. L2 glossing vs. no glossing	Reading: Receptive knowledge of 16 target L2 words	At least sixteen days	Favoured L1 use
S7	Arvizu (2020)	Hermosillo, Sonora, in North Western Mexico	Primary school EFL programme	Randomised controlled trial (with cluster randomisation)	167 Spanish L1 speakers, aged 9-12	Listening with the L1 storytelling method and pictorial method vs. no pictures or storytelling	Reading and listening: Receptive knowledge of 24 target L2 words	Just over five months	May have favoured L1 use, but not clear since multiple interventions were tested concurrently
S8	Baltova (1999)	Metropolitan Toronto, Ontario, Canada	Secondary school FSL programme	Randomised controlled trial (with cluster randomisation)	93 English speakers with a range of L1s	Watching with L1 audio and L2 captions vs. L2 audio and L2 captions vs. L2 audio only	Writing: Productive knowledge of 30 target L2 words	Around three weeks	Did not favour L1 use
S9	Baranowska (2020)	Słupca in the Greater Poland Voivodship in Poland	Secondary school EFL programme	Non-randomised comparison	63 Polish L1 speakers	Watching with L2 audio and L1 subtitles vs. L2 audio and L2 captions vs. L2	Writing: Receptive knowledge of 20 target L2 words	Around one hour	Did not favour L1 use

						audio only			
S10	Bueno-Alastuey and García (2019)	Pamplona, Spain	Secondary school EFL programme	Non-randomised comparison	52 Spanish L1 speakers, aged 12-15	The L1 translation method vs. the pictorial method vs. the video method	Reading and writing: Productive and receptive knowledge of 42 target L2 words	At least one week	Did not favour L1 use
S11	Busse et al. (2020)	An urban area of Lower Saxony in Northern Germany	Primary school EFL programme	Randomised controlled trial (with cluster randomisation)	42 German speakers, aged 8-10, with a range of L1s	Translation-based learning activities vs. textbook-based learning activities	Writing: Productive and receptive knowledge of 20 target L2 words	Just under two months	Favoured L1 use
S12	Busse et al. (2021)	Germany	Primary school EFL programme	Randomised controlled trial (with cluster randomisation)	51 German speakers, aged 8-10, with a range of L1s	Translation-based learning activities vs. motivation-based learning activities vs. textbook-based learning activities	Writing: Productive and receptive knowledge of 20 target L2 words	Just under four months (instead of just under two months, due to the coronavirus)	Favoured L1 use
S13	Camó and Ballester (2015)	Catalonia	Primary school EFL programme	Non-randomised comparison	34 Catalan L1 speakers, aged 10-11	Watching with L2 audio and target word pictorial flashcards with L1 oral translations vs. without L1	Listening: Receptive knowledge of 20 target L2 words	Around six weeks	No difference detected on the post-test, but favoured L1 use on both delayed post-tests

						translations			
S14	Chalmers (2019)	Oxford, United Kingdom	Primary school EAL programme	Randomised controlled trial (with individual randomisation)	40 EAL learners, aged 8-11, with a range of L1s	Watching vocabulary explanations in the L1 vs. in the L2	Writing and reading Productive and receptive knowledge of 18 target L2 words	Just over one month	No difference detected
S15	Choi et al. (2016)	A large metropolitan city in Korea	Secondary school EFL programme	Non-randomised comparison	180 Korean L1 speakers	Reading with L1 glossing vs. L2 glossing vs. no glossing	Writing: Receptive knowledge of 14 target L2 pseudowords	Five weeks	No difference detected on the post-test, but partly favoured L1 use on the delayed post-test
S16	Del Pilar Agustín Llach (2009)	Erkrath, Germany	Secondary school SFL programme	Randomised controlled trial (with individual randomisation)	27 German L1 speakers, aged 18-19	Reading with L1 glossing vs. writing sentences with L2 words following L1 translation vs. reading with a dictionary	Writing: Productive and receptive knowledge of 10 target L2 words	Three weeks	Favoured L1 use on both of the post-tests, but no difference on the delayed post-tests
S17	Emirmustafao-glu and Gökmen (2015)	Istanbul, Turkey	Primary school EFL programme	Non-randomised comparison	75 Turkish L1 speakers, aged 12 (on average)	The L1 translation method vs. the pictorial method	Writing: Productive knowledge of 20 target L2 words	At least five weeks	Did not favour L1 use on the post-tests, but no

									difference detected on the delayed post-tests
S18	Ghabanchi and Ayoubi (2012)	Iran	Secondary school EFL programme	Randomised controlled trial (with individual randomisation)	120 Persian L1 speakers	Reading with L1 glossing vs. L2 glossing vs. a dictionary vs. summary writing with target words	Writing: Receptive knowledge of 30 target L2 words	At least two weeks	No difference detected
S19	Gyllstad et al. (2023)	Urban areas in two parts of Sweden	Secondary school EFL programme	Non-randomised comparison	74 Swedish speakers, aged 14-16, with a range of L1s	Majority language definitions vs. majority and minority language definitions vs. L2-only definitions	Writing: Receptive knowledge of 60 target L2 words	Around four and a half months	Favoured L1 use on the post-test, but no difference detected on the delayed post-test
S20	He (2019)	China	Secondary school EFL programme	Randomised controlled trial (with individual randomisation)	6 Mandarin L1 speakers, aged 15.5 (on average)	Watching with L1 and L2 keyword captioning vs. L2 keyword captioning vs. no captioning	Speaking and writing: Productive knowledge of 31 target L2 keywords	Around an hour and a half	Favoured L1 use
S21	Helle (2014)	Austria and Norway	Secondary school EFL programme	Non-randomised comparison	123 Norwegian L1 speakers in Norway,	Watching with L2 audio and L1 subtitles vs. L2 audio only vs.	Reading: Receptive knowledge of	Not stated	No difference detected

					aged 14-16, and 89 German L1 speakers in Austria, also aged 14-16	L1 audio only	24 target L2 words		
S22	Hennebry et al. (2017)	The UK	Secondary school FFL programme	Randomised controlled trial (with cluster randomisation)	262 English L1 speakers, aged 14	Listening with L1 translations vs. L2 definitions or paraphrases vs. neither of these	Reading and writing: Receptive knowledge of 60 target L2 words	Twelve weeks	Favoured L1 use
S23	Hsu et al. (2013)	Taiwan	Primary school EFL programme	Non-randomised comparison	81 Chinese L1 speakers, aged 11 (on average)	Watching with L2 audio, L2 captions and L1 translations of the target words vs. L2 audio with both L2 captions and L1 translations of the target words only vs. L2 audio only	Reading: Receptive knowledge of 20 (or possibly 24) target L2 words	Just over one month	Favoured L1 use
S24	Hu et al. (2014)	Taiwan	Secondary school EFL programme	Non-randomised comparison	78 Chinese L1 speakers	Reading with optional L1 glossing vs. optional L2 glossing	Writing: Productive and receptive knowledge of 24 target L2 words	Four weeks	No difference detected on the post-tests, but did not favour L1

									use for high proficiency learners on one of the delayed post-tests
S25	Ina (2014)	Greece	Primary school IFL programme	Randomised controlled trial (with individual randomisation)	93 Greek speakers, aged 9-12	Watching with L2 audio and L1 subtitles vs. L2 audio only vs. L1 audio only	Listening and reading: Receptive knowledge of 35 target L2 words	Around one hour	Favoured L1 use
S26	Kongtawee and Sappapan (2018)	Thailand	Secondary school EFL programme	Non-randomised comparison	83 Thai L1 speakers	Reading with L1 hypertext glossing vs. L2 hypertext glossing	Reading: Receptive knowledge of 40 target L2 words	Around five weeks	Favoured L1 use
S27	Koolstra and Beentjes (1999)	An urban area of Rotterdam, Netherlands	Primary school EFL programme	Randomised controlled trial (with individual randomisation)	246 Dutch speakers, aged 9.7 years (on average) in Grade 4, and 11.7 years (on average) in Grade 6	Watching with L2 audio and L1 subtitles vs. L2 audio only vs. L1 audio only	Listening and reading: Receptive knowledge of 28 target L2 words	Around two weeks	Favoured L1 use
S28	Laufer and Girsai (2008a)	Israel	Secondary school EFL programme	Non-randomised comparison	48 Hebrew L1 speakers	Reading with contrastive analysis and translation activities vs.	Writing: Productive and receptive knowledge of 17 target L2	Just over two weeks	Favoured L1 use

						content discussion activities	words and phrases		
S29	Laufer and Girsai (2008b)	Israel	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	75 Hebrew L1 speakers, aged 15-16	Reading with contrastive analysis and translation activities vs. content discussion activities vs. vocabulary activities	Writing: Productive and receptive knowledge of 20 target L2 words and phrases	Around three and a half weeks	Favoured L1 use
S30	Laufer and Shmueli (1997)	Israel	Secondary school EFL programme	Non-randomised comparison	128 Hebrew L1 speakers	L1 translations and L2 definitions vs. L1 translations and L2 definitions with L2 example sentences vs. L1 translations and L2 definitions with a L2 example paragraph vs. L1 translations and L2 definitions with a L2 example paragraph and	Reading: Receptive knowledge of 20 target L2 words	Just over five weeks	Favoured L1 use

						in-text lexical elaboration vs. target L2 words only			
S31	Lee and Macaro (2013)	Incheon, Korea	Primary school EFL programme	Non-randomised comparison	443 Korean L1 speakers, aged 12	Reading and listening with L1 oral definitions vs. L2 oral definitions	Reading and writing: Receptive knowledge of 56 target L2 words	Nine weeks	Favoured L1 use
S32	López (2016)	Pamplona, Spain	Secondary school EFL programme	Non-randomised comparison	75 bilingual Basque and Spanish L1 speakers, aged 12-13	Listening with L1 translations vs. L2 definitions vs. neither of these	Writing and reading: Productive and receptive knowledge of 5 target L2 words	One week	Did not favour L1 use
S33	Lwo and Lin (2012)	Taiwan	Secondary school EFL programme	Non-randomised comparison	32 Mandarin L1 speakers, aged 14	Watching with L2 audio and L1 subtitles vs. L2 audio and L2 captions vs. L2 audio and both L1 and L2 captions vs. L2 audio only	Reading and writing: Receptive knowledge of 24 target L2 words	Five weeks	No difference detected
S34	Miyasako (2002)	Okayama, Japan	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	187 Japanese L1 speakers	Reading with L1 multiple-choice glossing vs. L2 multiple-choice glossing vs. L1	Reading: Receptive knowledge of 15 target L2 words	Eighteen days	Did not favour L1 use on the post-test, but no

						single glossing vs. L2 single glossing vs. no glossing vs. no reading			difference detected on the delayed post-test
S35	Muñoz et al. (2023)	Catalonia, Spain	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	39 Catalan and Spanish L1 speakers, aged 13.3 (on average)	Watching with L2 audio and L1 subtitles plus post listening activities vs. L2 audio and L2 captions plus post-listening activities	Writing: Receptive knowledge of 120 target L2 words	Eight months	No difference detected
S36	Peters et al. (2016) Experiment 1	Flanders, Belgium	General secondary school EFL programme	Non-randomised comparison	28 Dutch L1 speakers, aged 17-18	Watching with L2 audio and L1 subtitles vs. L2 audio and L2 captions	Writing: Receptive knowledge of 39 target L2 words	Just over one week	Did not favour L1 use
	Peters et al. (2016) Experiment 2	Flanders, Belgium	Vocational secondary school EFL programme	Non-randomised comparison	18 Dutch L1 speakers, aged 17-20	Watching with L2 audio and L1 subtitles vs. L2 audio and L2 captions	Writing and reading: Productive and receptive knowledge of 18 target L2 words	Just over one week	Did not favour L1 use
S37	Peters (2019)	Flanders, Belgium	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	118 Dutch L1 speakers, aged 16.4 (on average)	Watching with L2 audio and L1 subtitles vs. L2 audio and L2 captions vs. L2	Writing: Receptive knowledge of 36 target L2 words	Just over one week	Did not favour L1 use

						audio only			
S38	Pujadas and Muñoz (2019)	Barcelona, Spain	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	74 Catalan and Spanish L1 speakers, aged 13-14	Watching with L2 audio and L1 subtitles vs. L2 audio and L1 subtitles after a pre-listening task vs. L2 audio and L2 captions vs. L2 audio and L2 captions after a pre-listening task	Writing: Receptive knowledge of 120 target L2 words	Eight months	No difference detected
S39	Ramachandran and Rahim (2004)	Penang, Malaysia	Secondary school EFL programme	Non-randomised comparison	60 Malay L1 speakers, aged 16	Reading with L1 translations vs. L2 definitions	Writing: Receptive knowledge of 20 target L2 words	Four weeks	Favoured L1 use
S40	Sawada (2010)	Western Japan	Secondary school EFL programme	Non-randomised comparison	116 Japanese L1 speakers	Listening with L1 glossing vs. L2 glossing vs. no glossing	Writing and reading: Receptive knowledge of 45 target L2 words	Just under one month	Favoured L1 use on one of the post-tests and delayed post-tests, but no difference detected on the other

S41	Sieh (2008)	A suburban area of south Taiwan	Primary school EFL programme	Non-randomised comparison	63 Mandarin speakers, aged 9 years 5 months (on average)	Listening with target word pictorial flashcards and L1 oral translations vs. without L1 oral translations	Listening: Receptive knowledge of 52 target L2 words	Just over three weeks	No difference detected on Story A, but favoured L1 use on Story B
S42	Tonzar et al. (2009)	Treviso, Italy	Primary school EFL and GFL programmes	Non-randomised comparison	229 Italian L1 speakers, aged 9-13 on average	The L1 translation method vs. the pictorial method	Writing: Productive knowledge of 80 target L2 words (40 English and 40 German)	Around two months	Did not favour L1 use
S43	Wu and Yang (2022)	Rural China	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	133 Mandarin L1 speakers, aged 17-18	Watching with L2 audio and L2 captions with highlighted target words and L1 glosses vs. L2 audio and both L2 and L1 subtitles vs. L2 audio and both L2 and L1 subtitles with highlighted target words vs. L2 audio only	Writing and reading: Receptive knowledge of 10 target L2 words	Around 45 minutes	Favoured L1 use on two post-tests, but no difference detected on the third post-test

S44	Zhang and Graham (2020)	China	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	137 Chinese L1 speakers, aged 15-16	Listening with L1 definitions and example L2 sentences vs. L1 and L2 definitions with contrastive analysis vs. L2 definitions and example L2 sentences vs. no definitions	Writing: Receptive knowledge of 60 target L2 words	Twelve weeks	Favoured L1 use
S45	Zhang (2022)	China	Secondary school EFL programme	Randomised controlled trial (with cluster randomisation)	98 Chinese L1 speakers, aged 15-16	Listening with L1 definitions and example L2 sentences vs. L1 definitions with contrastive analysis vs. L2 definitions and example L2 sentences	Writing: Receptive knowledge of 20 target L2 words	Twelve weeks	Favoured L1 use

Key: ESL = English as a Second Language; EFL = English as a Foreign Language; EAL = English as an Additional Language;
FSL = French as a Second Language; FFL = French as a Foreign Language; GFL = German as a Foreign Language;
SFL = Spanish as a Foreign Language; IFL = Italian as a Foreign Language

Note: The 'study duration' section of this table does not include the time required for any pilot or preliminary studies to be conducted

Appendix 11: The EPHPP Risk of Bias Assessment Tool Guidelines

Quality Assessment Tool for Quantitative Studies Dictionary



The purpose of this dictionary is to describe items in the tool thereby assisting raters to score study quality. Due to under-reporting or lack of clarity in the primary study, raters will need to make judgements about the extent that bias may be present. When making judgements about each component, raters should form their opinion based upon information contained in the study rather than making inferences about what the authors intended.

A) SELECTION BIAS

(Q1) Participants are more likely to be representative of the target population if they are randomly selected from a comprehensive list of individuals in the target population (score very likely). They may not be representative if they are referred from a source (e.g. clinic) in a systematic manner (score somewhat likely) or self-referred (score not likely).

(Q2) Refers to the % of subjects in the control and intervention groups that agreed to participate in the study before they were assigned to intervention or control groups.

B) STUDY DESIGN

In this section, raters assess the likelihood of bias due to the allocation process in an experimental study. For observational studies, raters assess the extent that assessments of exposure and outcome are likely to be independent. Generally, the type of design is a good indicator of the extent of bias. In stronger designs, an equivalent control group is present and the allocation process is such that the investigators are unable to predict the sequence.

Randomized Controlled Trial (RCT)

An experimental design where investigators randomly allocate eligible people to an intervention or control group. A rater should describe a study as an RCT if the randomization sequence allows each study participant to have the same chance of receiving each intervention and the investigators could not predict which intervention was next. If the investigators do not describe the allocation process and only use the words 'random' or 'randomly', the study is described as a controlled clinical trial.

See below for more details.

Was the study described as randomized?

Score YES, if the authors used words such as random allocation, randomly assigned, and random assignment.

Score NO, if no mention of randomization is made.

Was the method of randomization described?

Score YES, if the authors describe any method used to generate a random allocation sequence.

Score NO, if the authors do not describe the allocation method or describe methods of allocation such as alternation, case record numbers, dates of birth, day of the week, and any allocation procedure that is entirely transparent before assignment, such as an open list of random numbers of assignments.
If NO is scored, then the study is a controlled clinical trial.

Was the method appropriate?

Score YES, if the randomization sequence allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which intervention was next. Examples of appropriate approaches include assignment of subjects by a central office unaware of subject characteristics, or sequentially numbered, sealed, opaque envelopes.

Score NO, if the randomization sequence is open to the individuals responsible for recruiting and allocating participants or providing the intervention, since those individuals can influence the allocation process, either knowingly or unknowingly.

If NO is scored, then the study is a controlled clinical trial.

Controlled Clinical Trial (CCT)
An experimental study design where the method of allocating study subjects to intervention or control groups is open to individuals responsible for recruiting subjects or providing the intervention. The method of allocation is transparent before assignment, e.g. an open list of random numbers or allocation by date of birth, etc.

Cohort analytic (two group pre and post)
An observational study design where groups are assembled according to whether or not exposure to the intervention has occurred. Exposure to the intervention is not under the control of the investigators. Study groups might be non-equivalent or not comparable on some feature that affects outcome.

Case control study
A retrospective study design where the investigators gather 'cases' of people who already have the outcome of interest and 'controls' who do not. Both groups are then questioned or their records examined about whether they received the intervention exposure of interest.

Cohort (one group pre + post (before and after)
The same group is pretested, given an intervention, and tested immediately after the intervention. The intervention group, by means of the pretest, act as their own control group.

Interrupted time series
A time series consists of multiple observations over time. Observations can be on the same units (e.g. individuals over time) or on different but similar units (e.g. student achievement scores for particular grade and school). Interrupted time series analysis requires knowing the specific point in the series when an intervention occurred.

C) CONFOUNDERS

By definition, a confounder is a variable that is associated with the intervention or exposure and causally related to the outcome of interest. Even in a robust study design, groups may not be balanced with respect to important variables prior to the intervention. The authors should indicate if confounders were controlled in the design (by stratification or matching) or in the analysis. If the allocation to intervention and control groups is randomized, the authors must report that the groups were balanced at baseline with respect to confounders (either in the text or a table).

D) BLINDING

(Q1) Assessors should be described as blinded to which participants were in the control and intervention groups. The purpose of blinding the outcome assessors (who might also be the care providers) is to protect against detection bias.

(Q2) Study participants should not be aware of (i.e. blinded to) the research question. The purpose of blinding the participants is to protect against reporting bias.

E) DATA COLLECTION METHODS

Tools for primary outcome measures must be described as reliable and valid. If 'face' validity or 'content' validity has been demonstrated, this is acceptable. Some sources from which data may be collected are described below:

Self reported data includes data that is collected from participants in the study (e.g. completing a questionnaire, survey, answering questions during an interview, etc.).

Assessment/Screening includes objective data that is retrieved by the researchers. (e.g. observations by investigators).

Medical Records/Vital Statistics refers to the types of formal records used for the extraction of the data.

Reliability and validity can be reported in the study or in a separate study. For example, some standard assessment tools have known reliability and validity.

F) WITHDRAWALS AND DROP-OUTS

Score **YES** if the authors describe BOTH the numbers and reasons for withdrawals and drop-outs.

Score **NO** if either the numbers or reasons for withdrawals and drop-outs are not reported.

The percentage of participants completing the study refers to the % of subjects remaining in the study at the final data collection period in all groups (i.e. control and intervention groups).

G) INTERVENTION INTEGRITY

The number of participants receiving the intended intervention should be noted (consider both frequency and intensity). For example, the authors may have reported that at least 80 percent of the participants received the complete intervention. The authors should describe a method of measuring if the intervention was provided to all participants the same way. As well, the authors should indicate if subjects received an unintended intervention that may have influenced the outcomes. For example, co-intervention occurs when the study group receives an additional intervention (other than that intended). In this case, it is possible that the effect of the intervention may be over-estimated. Contamination refers to situations where the control group accidentally receives the study intervention. This could result in an under-estimation of the impact of the intervention.

H) ANALYSIS APPROPRIATE TO QUESTION

Was the quantitative analysis appropriate to the research question being asked?

An intention-to-treat analysis is one in which all the participants in a trial are analyzed according to the intervention to which they were allocated, whether they received it or not. Intention-to-treat analyses are favoured in assessments of effectiveness as they mirror the noncompliance and treatment changes that are likely to occur when the intervention is used in practice, and because of the risk of attrition bias when participants are excluded from the analysis.

Component Ratings of Study:

For each of the six components A – F, use the following descriptions as a roadmap.

A) SELECTION BIAS

Strong: The selected individuals are very likely to be representative of the target population (Q1 is 1) **and** there is greater than 80% participation (Q2 is 1).

Moderate: The selected individuals are at least somewhat likely to be representative of the target population (Q1 is 1 or 2); **and** there is 60 - 79% participation (Q2 is 2). 'Moderate' may also be assigned if Q1 is 1 or 2 and Q2 is 5 (can't tell).

Weak: The selected individuals are not likely to be representative of the target population (Q1 is 3); **or** there is less than 60% participation (Q2 is 3) **or** selection is not described (Q1 is 4); and the level of participation is not described (Q2 is 5).

B) DESIGN

Strong: will be assigned to those articles that described RCTs and CCTs.

Moderate: will be assigned to those that described a cohort analytic study, a case control study, a cohort design, or an interrupted time series.

Weak: will be assigned to those that used any other method or did not state the method used.

C) CONFOUNDERS

Strong: will be assigned to those articles that controlled for at least 80% of relevant confounders (Q1 is 2); **or** (Q2 is 1).

Moderate: will be given to those studies that controlled for 60 – 79% of relevant confounders (Q1 is 1) **and** (Q2 is 2).

Weak: will be assigned when less than 60% of relevant confounders were controlled (Q1 is 1) **and** (Q2 is 3) **or** control of confounders was not described (Q1 is 3) **and** (Q2 is 4).

D) BLINDING

Strong: The outcome assessor is not aware of the intervention status of participants (Q1 is 2); **and t**he study participants are not aware of the research question (Q2 is 2).

Moderate: The outcome assessor is not aware of the intervention status of participants (Q1 is 2); **or** the study participants are not aware of the research question (Q2 is 2); **or b**linding is not described (Q1 is 3 and Q2 is 3).

Weak: The outcome assessor is aware of the intervention status of participants (Q1 is 1); **and** the study participants are aware of the research question (Q2 is 1).

E) DATA COLLECTION METHODS

Strong: The data collection tools have been shown to be valid (Q1 is 1); **and** the data collection tools have been shown to be reliable (Q2 is 1).

Moderate: The data collection tools have been shown to be valid (Q1 is 1); **and** the data collection tools have not been shown to be reliable (Q2 is 2) **or** reliability is not described (Q2 is 3).

Weak: The data collection tools have not been shown to be valid (Q1 is 2) **or** both reliability and validity are not described (Q1 is 3 and Q2 is 3).

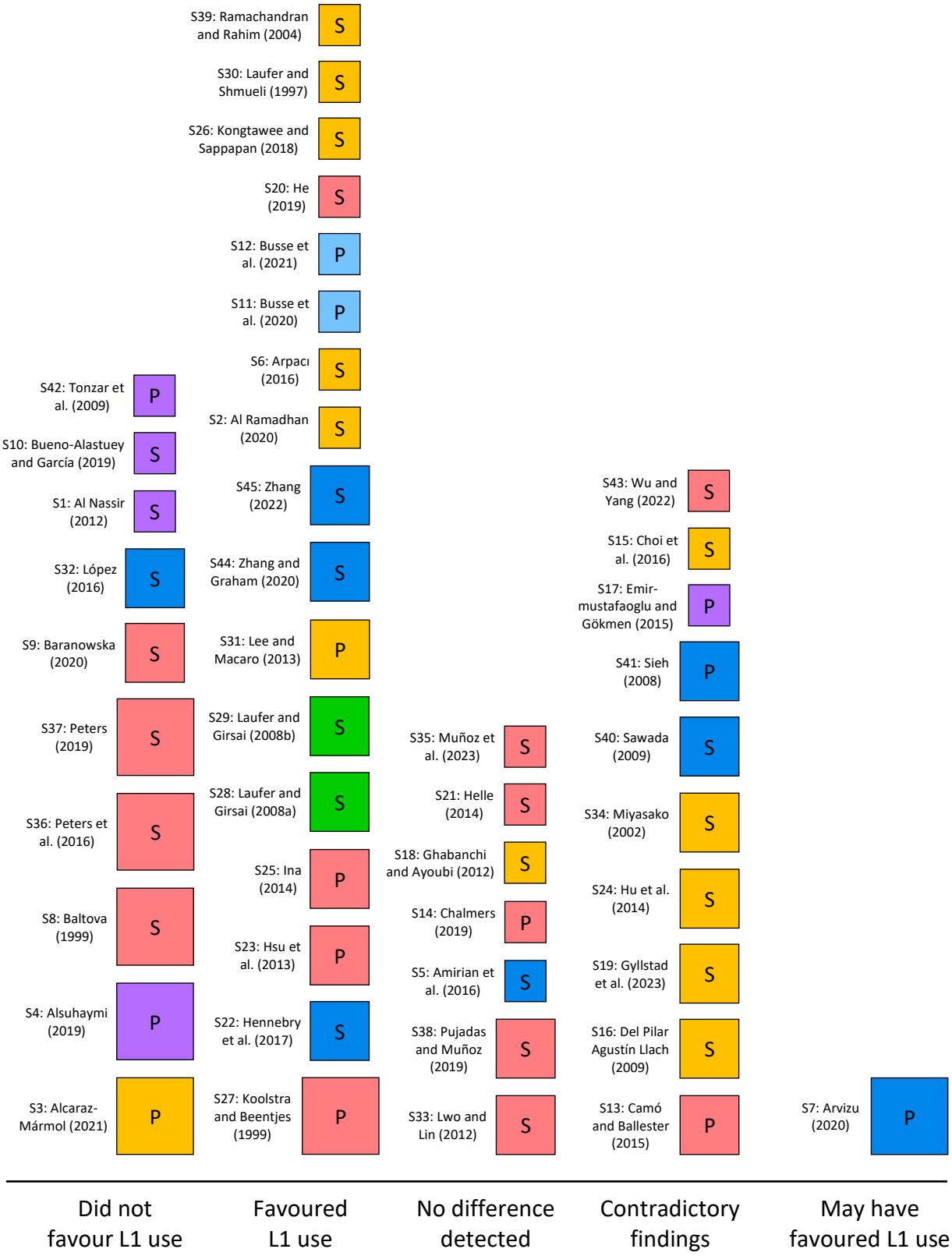
F) WITHDRAWALS AND DROP-OUTS - a rating of:

Strong: will be assigned when the follow-up rate is 80% or greater (Q2 is 1).

Moderate: will be assigned when the follow-up rate is 60 – 79% (Q2 is 2) **OR** Q2 is 5 (N/A).

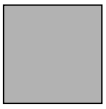
Weak: will be assigned when a follow-up rate is less than 60% (Q2 is 3) or if the withdrawals and drop-outs were not described (Q2 is 4).

Appendix 12: Study Outcomes in Relation to Intervention Type and Quality Assessment



Keys:

Risk of Bias Rating:



Strong rating
(Low risk of bias)



Moderate rating
(Moderate risk of bias)



Weak rating
(High risk of bias)

Educational Context:

S = Studies with secondary school-aged participants

P = Studies with primary school-aged participants

Vocabulary Teaching Intervention:



The L1 translation method vs. the pictorial/video method



Watching television with various combinations of L1/L2 audio, L1/L2 subtitles or oral definitions



Listening with various kinds of L1/L2 glossing, storytelling or contrastive analysis



Reading with various kinds of L1/L2 glossing or writing activities



Reading with content discussion, translation or vocabulary tasks



Translation-based learning activities vs. textbook-based learning activities