DEPOSIT AND CONSULTATION OF DISSERTATION

One copy of your dissertation will be deposited in ORA (Oxford University Research Archive), where it is intended to be freely available online. In order to facilitate this, you are requested to complete and sign the form below.

Please use block capitals

Surname **IQBAL**
First names (in full) **ASMA**
Faculty board **EDUCATION**
Degree name and pathway **M.Sc. Education: Child Development and Education**
Title of dissertation **Bilingualism and Mathematical Ability in Typically Developing Primary School Children: A Systematic Review** N.B. The title stated here must be precisely the same as that stated on the title page of the thesis submitted. **A candidate wishing to amend the title previously approved by the faculty must apply to the faculty board for permission to do so.**
Supervisor **Dr. Gillian West**
Subject keywords *Enter your own keywords or phrases to describe your work. This information helps us describe your work in ORA* **Bilingualism, bilingual education, mathematical ability, primary school children**
Research methods used *This information helps us describe your work on SOLO for future students* *e.g. quantitative, interviews, vocabulary test, systematic review, etc.* **Systematic review**

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 understand that the Department requires that I shall deposit one copy of my 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, 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 [www.bodleian.ox.ac.uk/ora/deposit-in-ora/deposit-licence] 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]:

Asma Iqbal

Date: August 12, 2022

Bilingualism and Mathematical Ability in Typically Developing Primary School Children: A Systematic Review

By

Asma Iqbal

Dissertation Submitted as Partial Fulfilment

of the Requirements for

M.Sc. Education

Pathway: Child Development and Education

Department of Education

University of Oxford, UK

August 2022

Supervisor: Dr. Gillian West

Abstract

A large body of research has been dedicated to studying cognitive advantages of bilingualism in all ages. However, the research investigating the link between bilingualism and mathematical attainment in young learners is limited. This systematic review aims to investigate the relationship between bilingualism and mathematical attainment in typically developing primary school children (aged 5-11) and examine the roles of moderators (age and gender) and mediators (SES and ethnicity) of bilingualism on mathematical ability. The main purpose of the review was to synthesize evidence based on the empirical research. The author identified the studies examining the relationship between bilingualism and mathematical ability of school aged children from 7 electronic databases. Of the 559 total studies identified, 23 matched the eligibility criteria. The author categorized the studies into 3 groups based on the programs or the experimental conditions that they studied. Mixed Methods Assessment Tool (2018) was used to assess the risk of bias of individual studies and a cumulative confidence in the studies was synthesized. The results show that two-third of the studies on bilingualism and bilingual education were conducted in the US. The overall results demonstrate that bilingualism is positively associated with mathematical ability. Important findings and implications for the future research have been discussed.

Acknowledgements

Special thanks to my Supervisor Dr. Gillian West whose invaluable knowledge of educational research and methodology, encouragement and sincere mentorship has enabled me to complete this project. Dr. Gillian has not only been my biggest support throughout this journey but also an inspiration to me. Furthermore, I am thankful to the staff at Education Library, especially Catherine Scut who gave me the required training for searching the databases and composing the Boolean search strings.

In addition, I am immensely grateful to my colleague, best friend, and mentor Alexis White, who has always been at my back in every matter of personal and professional life.

Special thanks to o my friends (Saeeda, Aqeela, Rabab, and Amber) and my beloved and soulmate Malik, and my Children Usman and Huda, whose love and motivating made me achieve my goals.

Lastly, my heartfelt thanks to Zarmina Azhar, my host in Oxford, who made me feel at home and has been so very generous and caring, that helped me focus on my studies peacefully.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the last decade, the number of refugees and migrants have doubled and exceeded 92 million worldwide, because of various political, climate change and economic reasons. Furthermore, Europe saw the largest displacement crisis in history, last year since the World War II. Due to their strong economies and political stability, Western countries are major destinations for the immigrants and refugees. According to the World Migration Report (2022), US was the largest immigrants' destination, receiving approximately 11 million Mexican immigrants in 2021. Based on the National Centre for Education Statistics report (2016), the percentage of the school aged immigrant and language minority children is expected to be 40% of the total school age population in the US by 2025. Canada, UK and Australia are no different from the US for the influx of the refugees and immigrants from all over the world.

The above scenario illuminates the phenomenal increase in the linguistic diversity of the primary school classrooms resulting from the displacement and calls for reconsiderations of the policy and practice for the use of native language in the classrooms. Prior research shows that there is no definite consensus on the educational programs to meet the needs of the immigrant and language minority students in majority English speaking countries. For instance, in the US and Canada, there are various government funded bilingual education programs in schools alongside the English-only programs, recognising the linguistic, academic, and sociocultural benefits of bilingual education. Conversely, in the UK, special education needs (SEN) services are offered to the new coming immigrant and refugee children, with little to no support in their native languages, indicating a discouragement of bilingualism and bilingual education. This implies that the debate between the proponents and opponents of bilingualism has not settled, and hundreds of thousands of language minority students' identities and opportunities to education are at the discretion of the decision makers.

Another important consideration in the primary and secondary education is mathematics learning which is crucial for students' success at school and opportunities to better career prospects (such as STEM careers) because math skills are required to succeed in all the science subjects. This implies that the schools must take necessary steps to make appropriate and practical provisions for the language minority students' mathematics achievement. To this end, bilingual education can be the right choice for these students who not only lose their linguistic identities in the English-only classrooms, but also fail to develop optimal fluency in both L1 and L2. Past research has established that bilingualism is associated with enhanced cognitive flexibility. If this holds true, the question arises that 'is bilingualism also linked with higher mathematical ability, since mathematics learning is related to general cognitive skills?' In the light of the promising results of the empirical research with

bilingual education and general academic achievement, there is need for investigating the relationship between bilingualism and mathematical ability in primary school children, in order to have them reap the benefits of bilingual proficiency, rather than taking away their linguistic identities in all-English educational programs. Hence, it would be worthwhile for the immigrant and language minority students to maintain their L1 while learning the language of instruction at school (L2).

To preserve the linguistic human rights of indigenous people and draw the global attention to the minority languages, the United Nations General Assembly (UNGA) has recognized the period between 2022 and 2032 as the International Decade of Indigenous Languages (IDIL 2022-2032). One of the UNGA's strategic action plan for the IDIL 2022-2032 aims is to direct academic research to devise appropriate methodologies to collect data, formulate felicitous language policies and find program and technological solution using Artificial Intelligence, to preserve and promote the indigenous languages world-wide. This invites the education researchers to make their contribution to the cause.

In the global context of language diversity, recognition of the indigenous languages by UNGA, and the linguistic, academic, and sociocultural challenges faced by the minority language students, the objective of this dissertation is to add to the body of knowledge in linguistic education research by exploring the relationship between bilingualism and mathematical ability, and synthesize evidence based on the empirical research to answer the above question. The findings of this dissertation will inform the policy and decision makers, future researchers, stakeholders, and educators to make rational decisions for the best interest of the minority language students' achievement in mathematics which will open doors for them to higher education and better careers.

Chapter 2

Review of Literature

2.1. Operational Definitions:

a. Bilingualism

Bilingualism refers to the ability to communicate in two languages, usually one's native language and an additional foreign language. However, bilinguals vary greatly with regards to their proficiency in L1 and L2, a phenomenon known as the 'degree of bilingualism' (Clarkson, 1992).

Prior research has shown that bilingualism and biliteracy, in educational contexts, is associated with higher academic performance in reading, language, and mathematics (Cummins, 1992). Thomas and Collier (2017) hold that bilingual students demonstrate high levels of cognitive development and increased engagement in the classrooms. Thomas and Collier (2017)'s longitudinal research on bilingual school has yielded positive results in favour of bilingual education (BE) which has impacted policy, practice and decision making in the department of education in the US such that bilingualism/multilingualism has become the most widely accepted pedagogy throughout the country, with 22 states recognising the biliteracy seal (Thomas & Collier, 2017).

b. Mathematical Ability

American Psychological Association (APA) defines Mathematical ability as students' capacity to solve quantitative problems including arithmetic and numeracy-related and word problems (n.d.).

c. Typically Developing Children

Typically developing children are the those who do not demonstrate any developmental, cognitive, learning or speech disorders and/or impairments.

d. Primary School

Primary school is the school for children between ages 5-11 (grades 1-6). In various education systems worldwide, the grade-levels in the primary or elementary schools differ in that in some countries primary school is considered from grade 1 to 5, while in others 6^{th} grade is also considered as primary school. Hence, in this dissertation, grade 1-6 are considered as primary school.

2.2 Research on Bilingualism and the Methodological Challenges

Research on bilingualism started in the 1910s with the development of intelligence tests and cognitive measures by the experimental psychologists (Hakuta, 1986). However, the research in this

field has been facing methodological challenges. Following section gives an overview of the previous research on bilingualism and its methodological challenges.

A. Historical Challenges

Hakuta (1986) points out that earlier research on bilingualism suffered from major methodological flaws and the investigators tended to make wrong interpretations of the results, which led to a debate on the cognitive advantages of bilingualism. More precisely, majority of the studies were on immigrants and viewed bilingualism as the cause of ‘language handicap’ or ‘mental confusion’ (Hakuta, 1986). For instance, Terman (1918) found that bilingual immigrant children's vocabulary and mental age scores (also assessed on a simple vocabulary test) correlated as well as those of native English students, after 3 to 4 years of bilingual education (Hakuta, 1986, p.24). On the contrary, Young (1922) argued in his paper that immigrant children were found to be linguistically 'handicapped' and performed below the American children in a 2-year longitudinal study. Later in his own dissertation, Young reported that the actual reason for the immigrant children's poor performance at school was that the teachers assessed their academic performance based on their capacity to use English (L2) and hence, verbal tests better predicted the students' school performance better than the non-verbal tests (Hakuta, 1986, p. 25-6). Moreover, Young clarified that the immigrants speaking non-English European languages performed better than did Asian immigrants speaking Japanese and Chinese that are very different from English language (Hakuta, 1986, p. 25-6). Likewise, Pintner (1923) found that the immigrant children performed below the national norms on the standardized intelligence test, but on the national norms on a researcher-developed non-language test of intelligence (Hakuta, 1986, p. 26). Mead (1927) reported that middle and high school students’ IQ scores associated with the amount of English language spoken at their homes and the number of years they had resided in the US (Hakuta, 1986, p. 26). Darsie (1926) compared the achievement of Japanese-American students on Stanford-Binet test to that of American students, and found that the Japanese students’ performance was poorer on the subtest involving more language component, which indicated that language ability was an important factor of academic achievement (Hakuta, 1986, p. 26). Similarly, Goodenough (1926) found negative correlation between the IQ level and the amount of non-English language spoken at home, and interpreted that some non-English ethnicities had average intelligence and were intellectually inferior (Hakuta, 1986, p. 27). In a study comparing bilingual students’ performance to that of the monolinguals, Smith (1923) found that monolinguals outperformed the bilinguals on dictation, sentence-completion and writing tasks in English. In a systematic study, Saer (1926) found that the rural bilinguals and monolinguals differed in their performance on language related subtests of Stanford-Binet, but reported no such differences between the urban bilinguals and monolinguals, and interpreted the findings on a psychodynamic basis (Hakuta, 1986, p. 29). Along the same lines, Yoshioka (1926)’ study with Japanese-American and native American students found negative impact of bilingualism on the students’ academic

achievement (Hakuta, 1986, p. 30). These studies and their findings illuminate two important points: (a) research on bilingualism continued to evolve in the first half of the 20th century and (b) the research suffered from major methodological flaws and the misinterpretation of the findings.

Hakuta (1986) holds that the research on bilingualism in later half of the 20th century attempted to address the methodological inadequacies. Peal and Lambert (1962)'s study, demonstrating positive results of bilingualism, was the most influential in this regard such that they drew a distinction between the 'balanced-bilinguals' and 'pseudo-linguals' and matched the sample for SES and the degree of bilingual proficiency. Peal and Lambert (1962) found that bilinguals demonstrated cognitive flexibility and outperformed monolinguals on both verbal and non-verbal measures of intelligence (Hakuta, 1986, p. 33-4). Ever since, peal and Lambert (1962) study, Hakuta holds that the researchers studying bilingualism in the latter half of the 20th century were inclined towards selecting a sample of 'balanced-bilinguals' and matching them for age, sex, SES, and any relevant variables, and finding positive results. This scenario led to a methodological bias instead of addressing the methodological flaws of the previous research in a meaningful manner (Hakuta, 1986, 35).

In his rigorously designed longitudinal study with extremely low SES Puerto Rican elementary school students in a bilingual education program, Hakuta (1986) found that the bilinguals' IQ scores positively correlated with their proficiency in L2 and their performance on the metalinguistic measures had positive association with their proficiency in L1. These findings suggested that bilingualism was more associated with non-verbal intelligence whereas proficiency in L1 was related to the metalinguistic awareness, and that longer exposure to bilingual education strengthened the students' proficiency skills in both languages was positively linked with their academic achievement (Hakuta, 1986, p. 39-41). These results also support Cummins (1979)'s threshold hypothesis that the bilingual students need to achieve sufficient proficiency in both languages to enjoy the cognitive advantages of bilingualism.

B. Current Challenges

A century has passed since the first study on bilingualism was reported (Terman, 1918 as cited in Hakuta, 1986), but the research in this area has not freed itself of the methodological challenges because the investigators have no control over some of the challenges. This holds true when the research involves children with low SES in linguistically diverse landscapes such as India and Africa where a multitude of languages is spoken in one area. The salient methodological challenges are as follows.

First, the bilingual students vary greatly in their proficiency level in both languages, which makes it difficult to identify an appropriate baseline group of monolinguals (Tsimpli et al., 2019). Second, research with disadvantaged low-income students poses a challenge for testing the students who are

not accustomed to formal conversations at school setting, however, this problem can be solved by recruiting and training local assistant researchers who are proficient in children's native language (Tsimpli et al., 2019). Third, many standardized assessments are not readily available in the minority languages, which is one of the reasons that very little research on bilingualism has been conducted outside the English-speaking countries. Moreover, the linguistic and cultural adaptations of the standardized tests may not be as valid and reliable as the original versions of the test (Tsimpli et al., 2019). Fourth, in low SES school settings, there are discrepancies between the official language of instruction and the actual language of instruction due to the lack of accountability and poor administration. This challenge may lead to disruptions in the research plans or encountering undesired circumstances for the research team regarding the choice of the language of the task for the students and/or the investigators' proficiency in the language of actual classroom instruction (Tsimpli et al., 2019). Fifth, longitudinal studies are considered a preferred research designs for bilingual research due to the nature of the phenomenon of bilingualism, but this design has a lot of challenges regarding the consistent use of the outcome measures throughout the entire duration of the study, students drop-out rate and attrition of the data, and change of the teaching and/or administrative staff members (Thomas & Collier, 2017). Sixth, random assignment of participants to the monolingual and bilingual groups is another big challenge (Willig, 1985). Lastly, it is difficult to control for all the student-level factors that may confound the results such family, home literacy environment, socio-cultural factors, parenting styles, children's use of technology devices in free time, involvement in leisure and sports activities and so on. These methodological flaws and challenges may lead to the decreased internal and external validity and generalizability of the research preventing the study findings from making meaningful contributions to the existing body of knowledge.

2.3. Theories of Second language acquisition

Early research on bilingualism attempted to find the effects of bilingualism, whereas later research was more focussed on finding the explanations of the bilingual advantage. Two important theories of second language acquisition have received the most attention in the research community in the field of bilingualism. These theories are discussed in the following section.

A. The Critical Period Hypothesis

The critical period hypothesis, first formulated by Penfield and Roberts (1959), explains the causes of the differences between the success of younger and older students' second language acquisition (www.psynso.com, n.d.). Lenneberg elaborated the critical period hypothesis stating that the age difference for second language acquisition is due to the 'brain maturation' such that the younger learners have an advantage over the older learners for developing proficiency in the second language. After the brain has gone through maturational changes (by puberty), learning a second language becomes hard. Hence, age intervenes the learning of L2 and cognition, and children learn

and develop native-like fluency in a foreign language faster than do adult learners (www.psynso.com, n.d.). However, Bialystok and Hakuta (1999) counter the claims made by the critical period hypothesis by emphasizing the interaction between the linguistic, cognitive, and socio-cultural factors that facilitate the second language acquisition. It is noteworthy that Bialystok and Hakuta (1999)'s refutations of the critical period hypothesis are in favour of Cummins (1979)'s theoretical framework of second language acquisition and his research on EAL students as discussed in the following section.

B. Cummins's Theoretical Framework: The Facilitation Theory

To address the previous researchers' contrasting explanations of bilingual children's poor academic performance, Cummins (1979) formulated the most influential theoretical framework to explain the minority-language students' second language acquisition and the patterns in their academic achievement in BE programs. This theoretical framework emphasizes that the 'interaction' between linguistic, socio-cultural, child-level and school-level factors plays a pivotal role in bilingual students' academic and cognitive development. Of the child-level factors, conceptual knowledge of L1, and the motivation to learn L2 and conserve proficiency in L1 play a mediating role between the socio-linguistic factors and BE program, and the child's academic achievement (Cummins, 1979). This interaction can be directly studies by assessing the child's level of competence in her native and second language. Hence, Cummins' theoretical framework of second language acquisition, placing a deep emphasis on the competency level in L1 and L2, proposed the 2 hypotheses to demystify the interrelations between language and cognition: (a) the threshold hypothesis and (b) the developmental interdependence hypothesis (Cummins, 1979).

a. The Threshold Hypothesis:

Cummins (1979)'s Threshold Hypothesis proposes that the children must attain an optimal level of proficiency in both languages (L1 and L2) to have the cognitive advantage of bilingualism demonstrated as higher academic achievement. Furthermore, Cummins points that there are two thresholds that are necessary for the child to achieve for having the bilingual advantage in their performance at school: the lower threshold (less than native-like proficiency) and the higher threshold (native-like proficiency). Children who fail to develop a minimum level of competence in L2 end up being semi-lingual, and even bi-semi-lingual on failing to achieve the competency in both L1 and L2 which in turn is linked with the poor academic achievement. Achieving the minimum threshold in L2 is essential to have no negative effects of bilingualism on academic achievement, whereas attainment of the higher threshold in L2 is associated with positive impact of bilingualism on academic and cognitive development. Though, the threshold hypothesis provides a theoretical framework to disentangle the link between different levels of bilingualism and academic achievement, it does not

clarify the link between L1 and L2 proficiency, which is explained by the second hypothesis (Cummins, 1979).

b. The Developmental Interdependence Hypothesis:

Cummins (1979)'s Developmental Interdependence Hypothesis (DIH), also called Common Underlying Proficiency (CUP) holds that the bilingual students' literacy skills in their two languages (L1 and L2) are interdependent in that the proficiency skills and academic knowledge transfer between the two languages, when the students receive sufficient support for both languages at school. More precisely, the L1 proficiency level defines the L2 proficiency level meaning that the stronger a child's L1 proficiency development is, the better the L2 proficiency development will be. On the contrary, children who have weaker L1 proficiency skills struggle to develop proficiency skills in L2, despite the increased exposure to L2 in the early school years. Furthermore, when the language of instruction is the students' first language, it not only facilitates their second language acquisition, and but also their academic development. To this end, for both minority and dominant- language students in bilingual education programs, instruction in L1 exerts no detrimental effects on students' achievement, despite the less exposure to L2 at school. In conclusion, the interaction between the language of instruction and the child's level of L1 competence before starting the school affect the L2 acquisition (Cummins, 1979). Skutnabb-Kangas and Toukoma (1978)'s UNESCO report on immigrant children also supports Cummins's developmental interdependence hypothesis such that immigrant Finnish children who had stronger L1 (Finnish) proficiency developed stronger proficiency in Swedish. In addition, Skutnabb-Kangas and Toukoma (1976) highlighted the need for strong L1 proficiency development for grasping abstract concepts in math and science subjects even when L2 was the language of instruction in these subject (Skutnabb-Kangas and Toukoma, 1976, as cited in Cummins, 1979).

The Two Dimensions of language Proficiency: BICS and CALP

Based on his extensive research on EAL students and their referral forms and psychological evaluations by their teachers, Jim Cummins further clarifies the two distinct dimension of language proficiency: basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP) (Skutnabb-Kangas, 2012). Specifically, BICS includes oral production of speech, pronunciation, and listening comprehension whereas CALP is associated with meta-linguistic skills that include reading comprehension, creative writing, and spelling accuracy. Cummins' work on EALs demonstrated that the bilingual students gained conversational proficiency in L2 earlier, but they took longer to develop academic proficiency in their second language (Skutnabb-Kangas, 2012). In other words, interpersonal communicative skills in L2 are developed in early childhood the cognitive skills in L2 are developed in late childhood when students have developed cognitive maturity, indicating a moderating role of age on bilingual students' second language acquisition

(Cummins, 1981). Cummins's re-evaluation of Ramsey and Wright (1974) data on English-as-a-second-language students in Canada, supported the notion that younger immigrants to Canada were at an advantage of developing BICS, whereas the older immigrants to Canada attained CALP rapidly (Cummins, 1981).

Additive Vs. Subtractive Bilingualism

Cummins (1976) argues that bilingualism is positively linked with divergent thinking (unstructured creative thinking that leads to novel and original solutions to the problems) and cognitive flexibility (ability to shift between mental tasks and strategies) in balanced bilinguals studying in BE programs that did not replace their L1 with L2 (additive bilingualism) who have attained sufficient proficiency in the 2 languages. On the contrary, bilingual students learning in transitional bilingual programs that replaces L1 with L2 gradually (subtractive bilingualism) demonstrate negative or no relationship with cognitive development. Cummins (1976) emphasizes the importance of balanced proficiency in the 2 languages in order to have a cognitive advantage. Hence, the level of competence in L1 and L2 (degree of bilingual proficiency) mediates the link between bilingualism and cognitive development. Cummins (1976) concludes that when the children have overcome the challenge of 'coping with the two languages' by attaining sufficient proficiency in both L1 and L2, bilingualism positively links to cognitive functioning.

An Interplay of Power through Language

Cummins (1986) goes on supporting bilingualism and advocates BE as a better pedagogical approach by looking at the interplay of the L1 and L2 from the lens of power at school and in broader social contexts such that the dominant language group exercises more power on the learners leaving very little power for the minority language group. In other words, Cummins (1986) addresses the 'unequal power relations' in the society and stretches the phenomenon of bilingualism from academic achievement and cognitive development to the students' identity negotiations with their teachers and the patterns of power in the classrooms. Cummins (1986) holds that in subtractive educational programs that suppress the students' native language by not acknowledging its worth and replacing it with L2, exercises 'coercive relations of power': power of the dominant over the weaker. In contrast, additive bilingualism, as Cummins points out, integrates the students' L1 in learning without undermining the development of the second language proficiency exerts 'collaborative relations of power' in the classroom where no group suppresses the other. To this end, additive bilingualism is a promising educational policy to close the achievement gaps between the dominant and the minority groups, and may, in turn, lead to the balance of power from the classrooms to the whole society. Cummins argues that the language minority students' academic achievement below the native English students reflects the unfair treatment of the school systems that assess the former on the standardized tests normed for the latter, and consequently label the indigenous students as academically

handicapped (Skutnabb-Kangas, 2012). In a nutshell, Cummins advocates the additive form of bilingual education if the goal of the school program is to develop the academic and cognitive development of the minority-language students to an optimal level (Cummins, 1979).

Despite its comprehensive explanations on second language acquisition, Rossell and Baker (1996) refute the claims made by Cummins (1979)'s Facilitation Theory, based on the flaws in the statistical analyses and interpretations of the findings of the studies that support the facilitation theory. In addition, Rossell and Baker (1996) highlight that the facilitation theory not only lacks rigorous empirical evidence, but also does not explain the reason of the sequence of learning L1 before L2. For instance, Rossell and Baker (1996) argue that the major drawback in Ramirez (1991)'s study that supported the facilitation theory was that the early- and late-exit programs were not statistically compared, rather the descriptive comparison were made using simple descriptive line graphs which led to the biased interpretations of the results in favour of the facilitation theory. Russell and Baker (1996) support the time-on-task hypothesis against Cummins (1979)'s facilitation theory.

C. Time-on-Task Hypothesis: English Only Model

Porter (1990) proposed that the best way to teach L2 was to expose the child to L2 at an early age (by 5), and the child should spend 'more time' effectively on learning L2 (cited by Cummins, 1992). The time-on-task hypothesis suggests that the second language acquisition takes place by spending more time on learning it, and dismisses the need for developing L1 proficiency (Rossell & Baker, 1996). To this end, Rossell and Baker (1996) claim that bilingual immersion and late-exit programs hold no superiority over the early-exit and partial immersion programs that provide L1 support only in the initial years of the program and transition to the full immersion model where all instruction is given in English only (or the target language).

In the light of the criticism on the Facilitation Theory, the importance of sociocultural aspects associated with the attainment of the L1 cannot be denied. As Cummins (1979) emphasizes, the transitional bilingual programs and the English only model suppress the minority language which raises issues of identity, equality, justice, and power in the society.

2.4. Bilingual Education Programs

Since the movement of bilingual education programs in the latter half of the 20th century, there have been attempts to implement and evaluate BE programs in the US to meet the ever-growing population of immigrant children to the country. In todays' world, the geographical mobility has increased immensely. According to National Centre for Education Statistics' report (2016), only in the US, the population of Hispanic, Asian, and American Indian students is expected to be 40% of the total school age population in the US by 2025 (Thomas & Collier, 2017). In this scenario, BE seems an urgent need to meet the educational and linguistic needs of the language-minority students. BE has

become the most popular, largely funded by the government, and widely endorsed educational pedagogy in the US indicating the impact of extensive research in this area. On the contrary, research on bilingualism is very little in the countries other than the US and so is the proportion of government funded BE programs and schools in those countries. Overall, the bilingual education programs can be grouped under 2 major categories based on the use of L1 and the goals of the bilingual education program: (a) additive bilingual education and (b) subtractive bilingual education (Cummins, 1979). However, there are also some other 'non-models' in practice for minority students in majority or foreign language contexts that force the minority child to learn through L2, and dismiss their linguistic rights and identity e.g., subtractive submersion education (Skutnabb-Kangas, 2012). These submersion programs are highly detrimental for the students' psychological, linguistic, and academic development (Skutnabb-Kangas, 2012).

A. Additive Bilingual Education

Additive bilingual education (ABE) programs give formal training in the child's both languages without replacing the L1 with L2. The goal of such programs is to make the students proficient in both languages (Cummins, 1979). There is a variety of ABE programs administered at elementary and secondary schools to meet the academic needs of the ELLs and LEPs. These include bilingual immersion, 2-way immersion, and dual language immersion programs, content and language integrated learning (CLIL) e.g., Spanish immersion program and Mandarin immersion program. ABE programs mostly follow a 90/10 model (90% instruction in the target language, 10% instruction in the partner language) in the beginning (K-1) transitioning to 50/50 model (equal time allocation to both languages) by the end of the primary school. In these programs, the partner language (English in most cases) is introduced by 3rd grade, and the core subjects are taught in the target language (Spanish/French in most cases) in the early years of primary school. As the students progress to the upper primary grades, they are expected to develop equal proficiency in both L1 and L2. Skutnabb-Kangas (2012) regards these ABE programs as strong models of bilingual education.

B. Subtractive Bilingual Education

Unlike ABE, subtractive bilingual education (SBE) programs, also regarded as 'mainstream education' gradually replace L1 with L2 (Skutnabb-Kangas, 2012). SBE programs expect the child to develop native-like proficiency in L2 and have little to no aims to maintain L1 proficiency (Cummins, 1979). Partial immersion, early- and late-exit, and 1-way immersion programs are the examples of subtractive bilingual education. Skutnabb-Kangas (2012) holds that The SBE programs are rather weaker models of bilingual education (Skutnabb-Kangas, 2012).

In addition to ABE and SBE programs discussed above, there also 2 more types of programs with the goal of preserving the L1 and teaching L2 as an additional or foreign language. A brief description of these programs is given below

Language Maintenance or Language Shelter Programs

Language maintenance or language shelter (LM/LS) programs teach through the medium of L1 by bilingual teachers, and L2 is taught as a second or foreign language. These programs exert no detrimental effects on child's L1 and academic achievement (Skutnabb-Kangas, 2012).

Language in Elementary Schools (FLES)

FLES programs involve teaching a foreign language to the students for a short period of time during regular school days. The variety of foreign languages offered in the FLES depends on the school type and its location. The program structure and objectives also vary ranging from 90 to 150 minutes of foreign language instruction per week. The curriculum and assessments also vary depending on the school type and its population. FLES programs have shown positive results by teaching students a foreign language at no cost of L1 proficiency and no loss of academic achievement.

2.5. Gaps in Research

Our discussion of the previous research in the beginning of this chapter indicates that myriads of research had been conducted to investigate the impact of bilingualism and BE on students' academic achievement and cognitive development. Most of these studies assess students reading, writing, and language proficiency and mathematics, putting greater emphasis on language-related outcomes, and have demonstrated contradicting results. Moreover, the research findings differ greatly based on the population, sample size, context/location of the study, age, SES, and cognitive level of the participants, and outcomes measures used. However, very little research has been done to isolate the relationship between bilingualism and students' mathematical ability.

Willing (1985)'s meta-analysis of selected studies on bilingual education found that when the tests were administered in English, the significant adjusted means for math were the lowest as compared with reading, language, and overall academic achievement. Moreover, the effect sizes were inconsistent for math. Willig (1985) takes account of a number of factors that may have yielded inconsistencies in the results including language of the test, effect size formula, academic domain of the criterion measures, BE programs, and the types of scores used in the analysis. Furthermore, Willig (1985) illuminates the substantial methodological challenges in research with BE that may lead to superficial findings.

Greene (1998) conducted a meta-analysis of 11 studies of good methodological rigor that investigated the effectiveness of bilingual education on students' academic achievement. Greene (1998) reported that the bilingual instruction improved students' performance in reading by .21 of a standard deviation, and .12 in math. Moreover, the gains for math were not significant (Greene, 1998).

These findings prompt 2 question: Does bilingual education have no advantage for mathematical achievement? Or why were the gains in math lower than those in English reading?

Thomas and Collier (2017)'s extensive research on bilingual schooling and academic achievement has presented positive results of bilingualism. However, their report does not isolate the effects of bilingual schooling on mathematical achievement.

In addition to the dearth of research with bilingualism and mathematical achievement in primary schools, no systematic review to date has examined the empirical work done in this field and synthesized the findings from the previous research. This systematic review aims to identify, appraise and synthesize empirical evidence to add to the existing body of knowledge in the field of bilingualism, BE and mathematical attainment of primary school students.

Chapter 3

Methodology

This research project deploys the methodology of systematic review. A systematic review is an analysis of the past research in the field, which is performed in "a systematic, rigorous and transparent manner" to address the gaps in and limitations of the evidence generated by the individual studies (Newman & Gough, 2019). Like health sciences, there has been an increased surge in the evidence-based movement in education over past few decades such as Education Endowment Fund in the UK funding the systematic literature reviews to examine the particular areas of interest and synthesize evidence to inform policy and decision making (www.educationendowmentfund.org.uk, n.d) and establishment of the DfEE Centre for Evidence-informed Policy and Practice in Education (EPPI-Centre). Thus, a systematic review is a form of secondary research aiming to answer a specific question by taking account of all the empirical evidence given by the previous body of primary research using a predefined eligibility criteria for inclusion of the past studies in the review, and synthesizes a cumulative evidence by analysing the findings and assessing the quality and strength of the evidence.

Before conducting this review, the author searched for other reviews in the field to avoid replication of the scholarly work conducted elsewhere, whether published or in progress. For this purpose, the author searched the PROSPERO database (The University of York, n.d.), the Campbell Collaboration Online Library (Campbell Collaboration, n.d), the EPPI Centre Database of Educational Research (EPPI-Centre, n.d.) and the International Database of Education Systematic Reviews (IDESR, n.d.) using the search terms 'bilingualism' and 'mathematical achievement'. The search retrieved only 1 result: an *ongoing* review in the PROSPERO titled "Language learning and mathematical achievement: A systematic review" (registration ID: CRD42020172859) which is partially relevant and does not meet the scope and criteria of the current review. In addition, to establish the accountability and transparency of the review, the author submitted the protocols on IDESR on 12th May 2020 before conducting the final search. The protocols were written following Moher et al. (2015)'s PRISMA-P guidelines, and registered with IDESR on 12th may, 2022. A complete description of the protocols registered with IDESR is given in the appendix A. Detailed eligibility criteria and methodology follows in the next section.

3.1. Review Questions

Articulating specific review questions is an essential feature of conducting a systematic review as it guides the researcher to design the particular structure of the review and make vital decisions throughout the research process (Newman & Gough, 2019). This systematic review seeks to answer the following questions.

A. Primary Question

Q1. *Q1. Do the typically developing bilingual children in primary schools have higher mathematical ability as compared with the monolinguals?*

B. Secondary Questions

Q2: Do the typically developing primary school students attending bilingual education (BE), transitional bilingual education (TBE) or foreign language in elementary school (FLES) programs demonstrate higher mathematical ability than monolingual students in English only (EO) programs?

Q3: At what age/grade level, there is consistent math advantage of bilingualism?

Q4. Do age and gender of the leaners have moderating effects?

Q5. Do the SES and ethnicity have mediating effects?

Q6. Do bilingual assessments (dual language tests or test items) affect children's performance and achievement in math?

Q7. Why are there inconsistent finding in the studies with bilingualism and mathematical ability?

3.2. Eligibility Criteria

Table 3.1 presents the eligibility criteria for the inclusion of the studies in this review based on the bibliographic information, population, outcome measures, methodology, and language. However, publication date and publication type were not included in the eligibility criteria to broaden the scope of the review.

Table 3.1: Eligibility Criteria

Item	Include	Exclude	Rationale
Bibliographic Information	1. Open access studies that contained complete information about the author, and the publication date and type	1. Studies that lacked sufficient bibliographic information and were not open access	Sufficient bibliographic information is necessary to retrieve the full text.
Population	2. Studies with typically developing primary school students (TDPSS) (grades 1-6)	2a. Studies that explicitly target or include children who have any sort of learning or speech disorders and/or cognitive impairments 2b. Studies with pre-schoolers, middle and high school students, and adults	The review aims to examine the link between bilingualism on TDPSS's mathematical ability (MA). As clarified earlier, this is the first review of its kind. Hence, the author has narrowed down its scope to typically developing children to synthesize cumulative evidence of the research in this field with this specific population. Moreover, typically developing children differ greatly from non-typically developing children with regards to the cognitive abilities, which makes the two groups non-comparable for the bilingual advantage manifested in MA.
Outcome measures	3a. Studies that have language related outcomes (e.g., reading, vocabulary, spelling, grammar etc.) or language has been used as a predictor. 3b. Studies that have outcomes related to MA in any form (word problems, verbal/non-verbal arithmetic skills, numeracy, or mixed math skills).	3. Studies that do not include a measure of MA 3b. Studies involving sign language and sign bilingualism, brain studies. non-bilingual context.	To study the link between bilingualism and MA, there must be language and math related outcomes for the investigators to find empirical evidence using quantitative data.
Methodology: a. Study design	4a. Longitudinal, cross-sectional, correlational, intervention, small scale and pilot studies that are primary/empirical research	4a. Qualitative studies (case studies, observational studies) and systematic reviews 4b. Studies that don't involve a bilingual context in the classrooms. Studies	a. Qualitative studies are not truly systematic, and hence, have not been included in this systematic review. b. Pertaining to the purpose of the review, bilingualism, or BE is the key variable that

b. Intervention or experimental conditions c. Control group	4b. Studies that involve bilingual instruction in any form and measure MA. 4c. Presence of an appropriate control group of monolinguals or students in monolingual educational program or language effect is controlled statistically.	done with bilingual tech tools or software and programs as these are short term focus of bilingual literacy and can't be assessed empirically due to various factors including time spent on the software (e.g., Duolingo) by each child, reliability, and validity of the software or web portals etc. 4c. Studies that neither include a control group nor control for the language effect the statistical analyses.	must have been investigated in the study. c. Without an appropriate control group of participants, the study lacks strong internal validity, and the results cannot be attributed to the independent variable with confidence.
Language	5. Studies published in English	5. Studies not published in English	Studies that are available in full text in English will only be included in this review as Google translator is not a reliable tool for assessing the quality of and interpreting the findings of scientific papers.

3.3. Search Strategy

Following section gives a detailed account of the search strategy deployed in this review.

a. Information Sources

A total of 7 eminent electronic databases in the field of education, psychology, linguistics, and multidisciplinary research were searched to retrieve the studies for this review via Oxford University's subscription. The databases include EBSCO host, PsycINFO, MEDLINE, SCOPUS, Web of science Core Collection, ProQuest Social Science Premium Collection, and ProQuest Dissertation & Theses Global.

b. Steps to Improve the Search Strategy

The author took following steps to improve the search strategy.

i. Training with the Research Librarian: Prior to conducting this systematic review, the author collaborated with the research librarians at Oxford University's Department of Education to get necessary training and guidance to search the electronic databases, enhance the search strategy and develop Boolean search strings.

ii. Pilot Search: The author did a pilot search to select a range of appropriate keywords picked from relevant articles and develop the search terms.
iii. Primary Search: The author searched the 7 electronic databases modifying the Boolean search for each database accordingly.
iv. Supplementary Search: Next, the author hand searched key journals retrieved in the pilot and primary search that may have been missed in the primary search. This strategy of supplementing the search is efficient because sometimes the keywords have not been indexed on the databases because of using different synonyms. Additionally, the author searched Google scholar using the two key terms: 'bilingualism' and 'mathematical ability'.
v. Citation Checking: A bunch of relevant studies were retrieved on checking the citations found at 'full-text screening' stage. These studies were escaped in the primary search.
vi. Searching Grey or Fugitive Literature: Grey literature consists of documents that include final reports required by funding agencies, technical reports, program evaluation reports, and discussion papers, undergraduate and postgraduate theses, and conference proceedings. To make sure that this review does not skip any relevant studies, Oxford Research Archive (ORA) and Bodleian Libraries were also searched.

c. Search Terms

An efficient search strategy for conducting a systematic review is constructed by including the terms related to the intervention or the IV, outcomes, and population of interest, and combining them by using Boolean operators 'AND' and 'OR'(Newman & Gough, 2020), The author developed a comprehensive search strategy by including the keywords determined in the pilot search that started with 'bilingualism', 'mathematics', 'ability', and primary school', and returned a small number of studies. The authors, then refined the key terms by reviewing the results retrieved in the pilot search and included further terms to make the search strategy more efficient (table 4.2).

Table 4.2. Boolean Search Terms

Bilingualism OR "foreign language learning" OR "English language learning" OR "English as a second language"	AND	math OR mathematics OR arithmetic	AND	Success OR achievement OR attainment OR ability	AND	"Primary school students" OR "elementary school students" OR "grade 1" OR "grade 2" OR "grade 3" or "grade 4" Or "grade 5" OR "grade 6"

3.4. Data Management

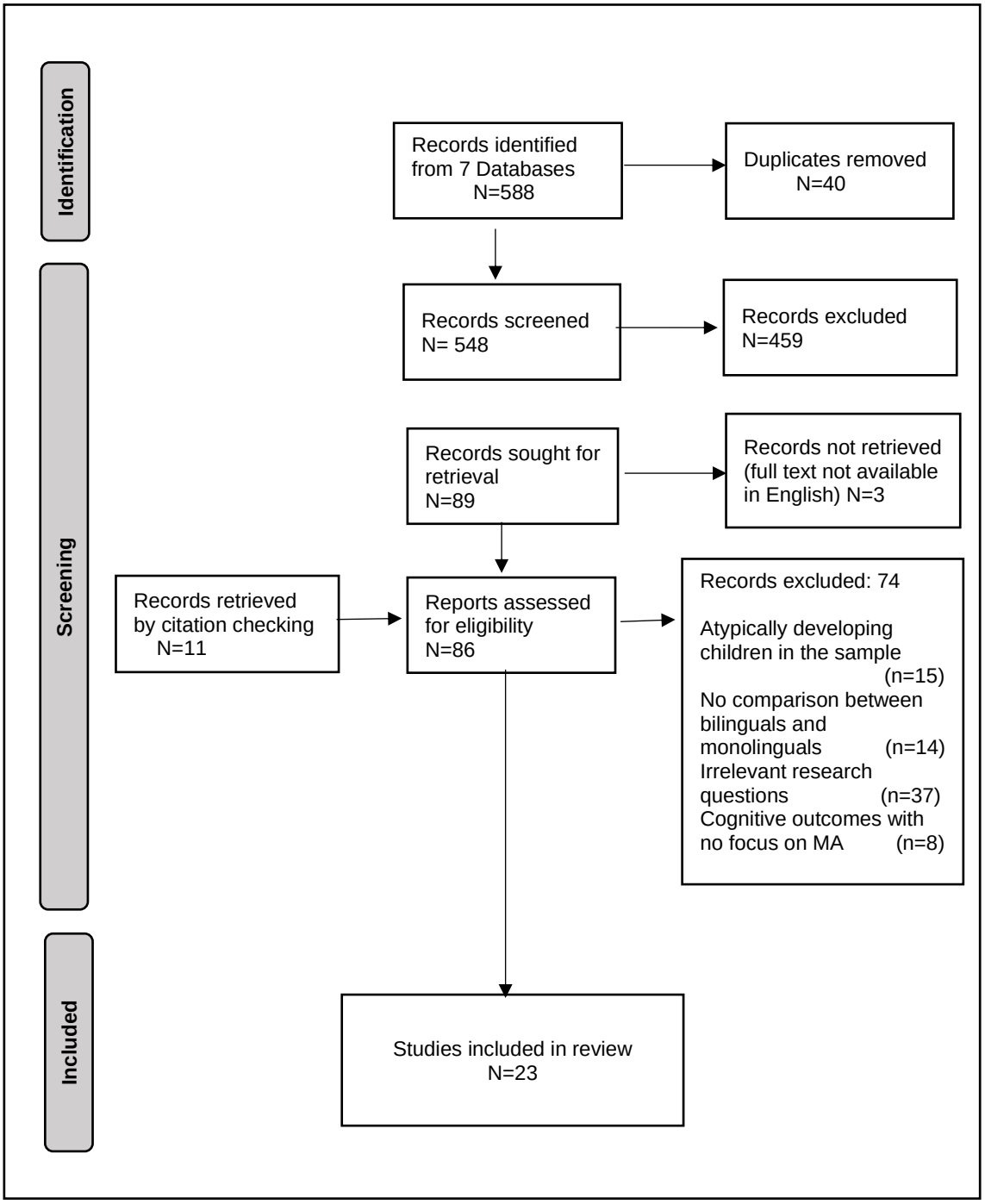
The search results from all the electronic databases were exported to EndNote in RIS format, and duplicates were removed. To keep a track of the shortlisted studies from each database, the author

created separate folders in her EndNote library for each database. Next, the references were exported to Rayyan data management software for initial screening. Rayyan also detected some duplicates which were removed on further screening. Moreover, the references in Rayyan were blinded to both the main author and the collaborator to calculate the reliability of the decisions at initial screening. After the initial screening, the author downloaded full texts of all the shortlisted studies and organized them in her computer for the full-text screening in Adobe Acrobat Reader DC (64-bit) that allows the users to highlight and annotate important parts of the text. During the entire process of screening and data extraction, the author also maintained a digital journal to document the search process and the key findings and unique features of each study. For data extraction, the author created a data extraction spreadsheet in Microsoft Excel. All the research was done on Huawei MateBook D 14.

3.5. The Process of Study Selection

The study selection process in a systematic review is executed in two steps by first screening the study titles abstracts and next, screening the full text (Newman & Gough, 2019). A description of the screening process is given below. Figure 1 shows the flowchart with a breakdown of the studies in the identification and screening process.

Figure 3.1. Flowchart for Study Selection



After importing the references from EndNote to Rayyan, the author invited her dissertation supervisor/collaboartor to the review on Rayyan. The decisions on screening were blinded to each other. The primary search returned 588 results which were narrowed down to 548 after removing duplicates in EndNote and Rayyan. In the initial screening, titles and abstracts were screened following a pre-defined criterion: (a) the study is quantitative research; (b) the study involves

typically developing primary school children; (c) the study includes measures related to language ability and mathematics achievement; and (d) there is an experimental group of bilinguals or students studying in bilingual education programs. If the answer was 'No' to any of the above criteria, the study was excluded. The initial screening resulted in 79 studies with the 'include' decision and 10 studies with 'May be' decision due to insufficient information in the abstract. In total, 459 studies did not meet the initial screening criteria due to wrong outcomes (163), wrong population (105), irrelevant outcomes (16), full text available in German only (3), and having no bilingual context (129). Additionally, some excluded studies were non-empirical studies (5), systematic reviews (3), meta-analysis (1), discourse analysis (9), observational study (4) and included no math related outcomes (45).

i. Full-text Screening

For the full-text screening, the author retrieved the full texts and manually screened each study for the relevance, adding 2 more criteria to initial screening.

a. Does the study include group means and standard deviations to enable a measure of difference between groups to be calculated? Or does the study deploy a method of inferential statistics to draw a meaningful conclusion? Alternatively, does it include a measure of effect size (r) for the relationship between MA and language in correlational studies?
b. Does the study include an appropriate control group of monolinguals or the students studying in monolingual programs? Alternatively, does the study control for language in the statistical analyses?

If the answer was 'No' to any of these questions, the study was excluded from the review. Only studies that fully matched the eligibility criteria were retained for data extraction. Studies were excluded at full text screening for lacking appropriate control groups, including atypically developing children in the sample, and not clearly investigating the link between bilingualism and MA. Of the 79 studies shortlisted at initial screening, 18 passed the full-test screening criteria, and out of the 10 studies labelled as 'May be', 2 were included in the review. During full text screening, 'citation checking' retrieved another 11 studies. A total of 23 studies were shortlisted for data extraction.

ii. Interrater Reliability for the Screening Process

The main author also double checked the abstracts in Rayyan. To ensure the reliability of the main author's screening decisions, a collaborator well-versed in education research reviewed 12% of the abstracts. There was 0.81 interrater reliability between the main author and the collaborator on initial screening decisions. However, it was not possible for the second reader to review the full texts. Hence, the main author discussed her decisions on full texts with her supervisor in detail. The

interrater reliability coefficient indicates that there was low risk of the systematic bias in the study selection process and the eligibility criteria had been appropriately applied.

3.6. Data Extraction

For data extraction, the author designed a data extraction form in Microsoft word, following the guidelines of Cochrane Data Extraction form for RCTs and non-RCTs, and piloted the form on 3 studies. Necessary adjustments were made to the form to make it suitable for all types of studies included in the review. The form included necessary categories for extracting important data about participants, intervention, comparison, and outcomes. Next, the data was recorded on Excel spreadsheet. A final version of the data extraction form is given in appendix C.

3.7. Risk of Bias Assessment of Individual Studies

In a systematic review, it is crucial to critically appraise the studies for their weaknesses and strengths and assess the risk of bias in the studies to determine the level of confidence in the studies and findings of the review (Hong & Pluye, 2018). For the risk of bias assessment (RoBA), the Mixed Methods Appraisal Tool Version 2018 developed by Pluye et al. (2018). As the studies that passed the eligibility criteria were all quantitative non-randomized studies, the 3rd section of the MMAT (2018) "Quantitative non-randomized" was used for the critical appraisal of the studies (Appendix C). The studies were rated for the risk of bias against the 5 descriptors on the MMAT 92018), and a global weight of the evidence rating was given to each study.

3.8. Data Synthesis

Due to the non-comparability of the studies based on the different contexts and experimental conditions, a meta-analysis was not possible. Hence, a narrative synthesis of the data was made after grouping the studies into logical categories, assessing the risk of bias, and synthesizing the findings of the groups of the studies.

Chapter 4

Results

This chapter presents results of the review. The chapter is organized into 4 sections. First section presents the general characteristics and overall statistics and trends in the included studies, and the second section gives brief accounts of the included studies. The risk of bias assessment of the individual studies forms the third section which is followed by the fourth section that answers the primary and secondary questions asked in this review.

4.1. General Characteristics

Table 4.1 presents general characteristics of the studies included in the present review.

Table 4.1. Characteristics of the Included Studies

Study Reference	Publication Type	Study Location	Study Design	Educational Program Tested (IV)	Sample Age-group	Sample Size	Study Duration	Control Group	Outcome measure(s)	Results	Effect Size
1. Lopato, 1963	Journal Article	US	Q-E[1]	FLES	G3	114	1 year	Yes	Standardized	Positive	Not given
2. McConnell, 1980	Doctoral Dissertation	US	Q-E	IBI	PreK - G3	1,020	6 years	Baseline model	Standardized	Positive	Not given
3. Clarkson, 1992	Journal Article	Papua New Guinea	Q-E	Bilingualism	G3	301	N/A	Yes	RD*	Positive	Not given
4. Kempert et al., 2011	Journal Article	Germany	Q-E	Bilingualism	G3	78	N/A	Yes	RD*	Positive	Small
5. Padilla et al., 2013	Journal Article	US	Q-E	BIP	G2-G5	1,029	4 years	Yes	Standardized	Positive	Not given
6. Lee et al., 2015	Journal Article	Cambodia	Q-E	TBE	G1-G4	100	4 years	Yes	RD*	Positive	Large
7. Fayt, 2019	Master's Dissertation	Netherlands	Q-E	BE	G2	76	N/A	Yes	RD*	Positive	Medium
8. Fleckenstein, 2019	Journal Article	Germany	Q-E	PIP	G1-G4	590	4 years	Yes	Standardized	Positive	Not given
9. Barbu et al., 2019	Journal Article	France	Q-E	Foreign CLIL	G1	116	1 year	Yes	RD*	No significant relationship	Zero
10. Rafferty, 1985	Research Report	US	C-EPF[1]	FLES	G3-G5	13,200	3 years	Yes	Standardized	Positive	Not given
11. Lee, 1989	Journal Article	US	C-EPF	BE	G6-G6	544	N/A	Yes	Standardized	Positive	Not given
12. Saunders, 1998	Doctoral Dissertation	US	C-EPF	FLES	G3	822	1 year	Yes	Standardized	Positive	Not given

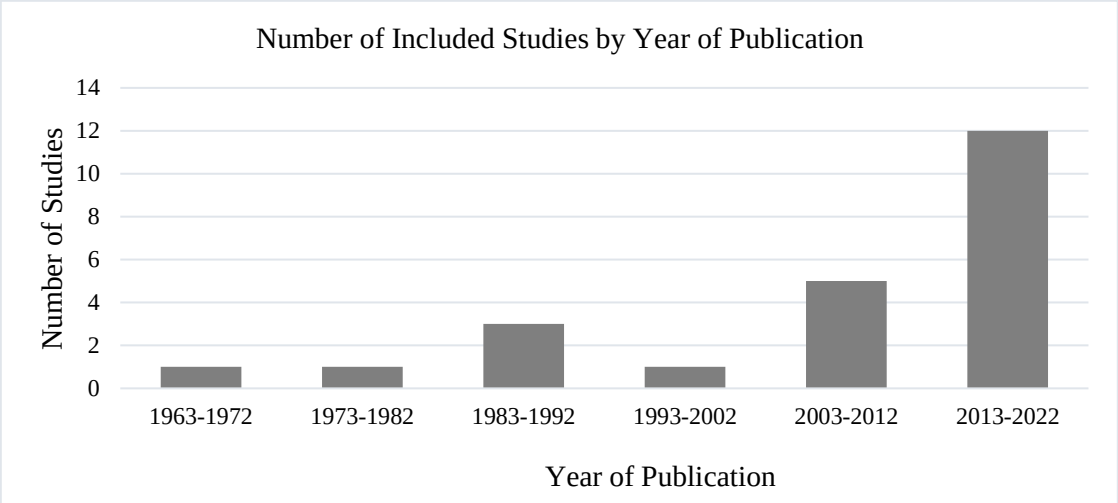
13. Leon, 2006	Doctoral Dissertation	US	L-EPF[3]	SPIP	G3-G5	171	3 years	Yes	Standardized	Positive	Moderate to large
14. Sanders, 2010	Doctoral Dissertation	US	L-EPF	BIP	G3-G5	846	3 years	Yes	Standardized	Positive	Not Given
15. Taylor and Lafayette, 2010	Journal Article	US	L-EPF	FLES	G3-G5	4,345	3 years	Yes	Standardized	Positive	Near zero (trivial)
16. Han, 2012	Journal Article	US	L-EPF		K-5	16,380	6 years	Yes	Standardized	Positive	Not given
17. Broomes, 2013	Journal Article	Canada	L-EPF	Bilingualism	G3, G6	121,037	N/A	Yes	Standardized	Positive	Not given
18. Fuhriman-Cleverly, 2014	Doctoral Dissertation	US	L-EPF	DLI	G5	106	N/A	Yes	Standardized	Positive (No significant differences)	Not Given
19. Rega, 2015	Doctoral Dissertation	US	L-EPF	SIP	G5	80	6 years	Yes	Standardized	Positive (No significant differences)	Not significant
20. Choy, 2016	Master's Dissertation	US	L-EPF	Bilingualism	K-5	8,036	6 years	Yes	Standardized	Positive	Small to medium (R^2)
21. Steele et al., 2017	Journal Article	US	L-EPF	DLI	K-8	1,625	9 years	Yes	Standardized	No significant relationship	Not given
22. Hartanto et al., 2018	Journal Article	US	L-EPF	Bilingualism	G1	13,439	1 year	Yes	Standardized	Positive	Not given
23. Serafini et al., 2020	Journal Article	US	L-EPF	TBE	K-5	20,870	6 years	Statistical controls	Standardized	Positive	Not given

[1]Quasi-experimental Design; [2]Crossectional Ex Post Facto Research; [3]Longitudinal Ex Post Facto Research; *Researcher-Developed

4.1.1. Publication Date

Research in bilingualism and mathematical ability (MA) spans over 6 decades and dates to early 1960s when the movement for bilingual education started in the US. However, over the last decade, the research has prominently increased in this field. Figure 4.1 shows the number of the studies by publication year, included in the present review.

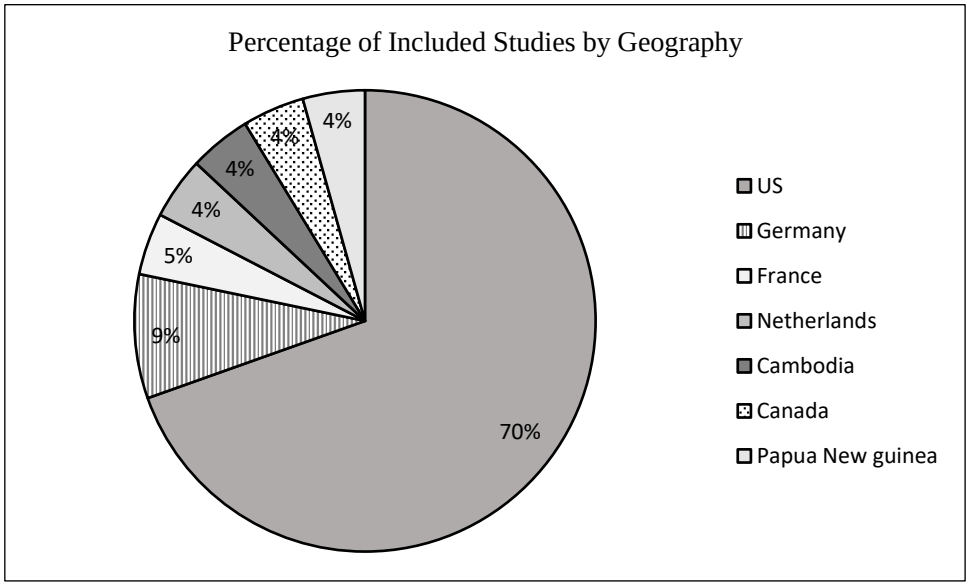
Figure 4.1: Number of Studies by Publication Year



4.1.2. Geography

Figure 2 demonstrate that a high proportion of the research studying the relationship between bilingualism and MA has been conducted in the US since 1963.

Figure 4.2: Studies by Geography



4.1.3. Publication Type

Of the 23 studies in this review, 5 are doctoral dissertations, 2 master's dissertations, 1 research report, and 15 journal articles.

4.1.4. Study Design

Results show that quasi-experimental design and ex post facto (causal-comparative) design are the two types of research designs that have been deployed in quantitative research investigating the link between bilingualism and mathematical ability in TDPSS. Specifically, 39% of the studies are quasi-experimental lacking the random assignment of the participants to the experimental and control groups whereas 61% of the studies are causal-comparative that retrieved students' assessment data from the past to examine the two constructs.

Quasi-Experimental Research:

Quasi-experimental research is a type of scientific research that attempts to establish cause-and-effect relationship between the predictor and outcome variables by using an experimental and a control group, without random assignment of the participants to the control and experimental groups due to the ethical or practical issues (SAGE Research Methods, 2012). Although not purely scientific, carefully designed quasi-experiments can allow to establish causality (SAGE Research Methods, 2012).

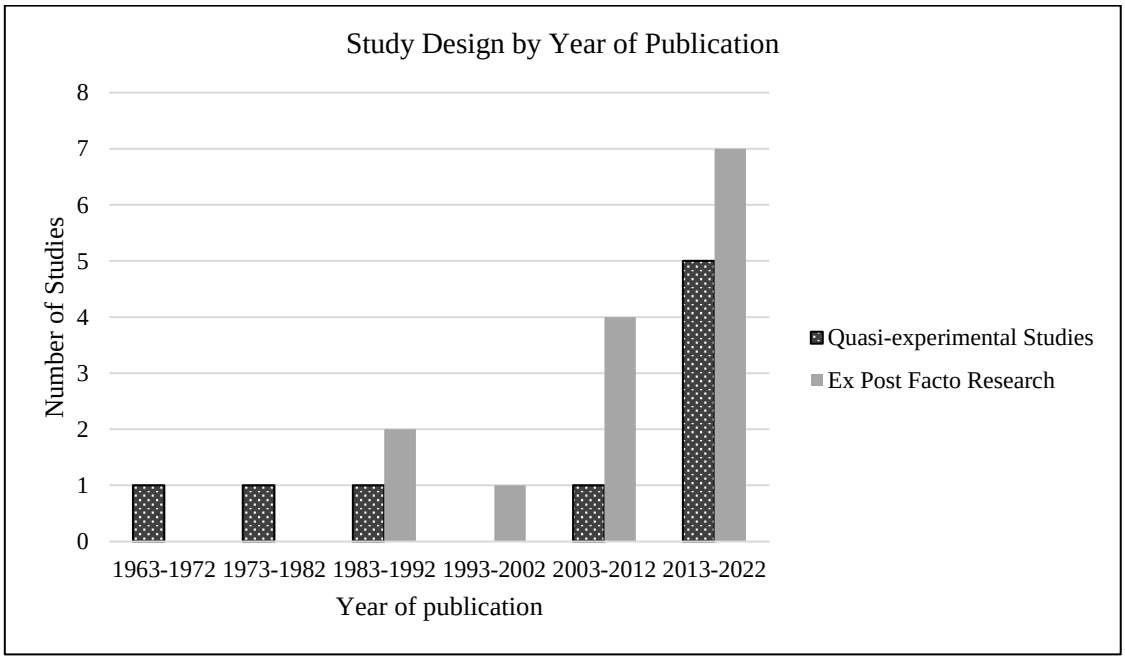
Ex post facto design:

Ex post facto study or after-the-fact research is a category of experimental research design in which the researcher can no longer manipulate the predictor variable because the outcomes have already occurred (SAGE Research Methods, 2012). Due its fundamental logic of investigation, ex post facto research is considered a category of experimental research design and is used as an alternative of the true experiments to test the hypotheses for establishing causality. Majority of the research in social sciences is ex post facto research (SAGE Research Methods, 2012). In education, ex post facto research retrieves data from (e.g., assessment data) the past to understand the trends in the phenomenon (e.g., academic achievement) in the past and get a perspective on the present and future.

Figure 3 shows that there emerged a trend of conducting ex post facto research since in 1980s by analysing longitudinal state-wide or district assessment data to find the students' academic growth patterns and evaluation of the government funded bilingual education programs in primary schools. Most of these ex post facto studies are large scale studies. It is noteworthy that a two-third of the included studies conducted in the US were ex post fact design whereas a one-third were quasi experiments. In contrast, in the countries outside the US, the research on bilingualism and MA, with

TDPSS, was entirely quasi-experimental except for Canada where ex post facto design was adopted in this research area.

Figure 4.3: Study Design by Year of Publication



4.1.5. Logical Categories of the studies

Table 4.2 shows the classification of the included studies into logical categories, indicating that 4 of the studies investigated the impact of FLES programs on students' MA, 6 examined the link between bilingual proficiency and MA, and 13 explored the effect of BE on MA.

Table 4.2. Logical Categories of the Studies

Study Designs	Interventions and Conditions Tested		
	FLES[1] Programs	**BE[2] Programs**	**BP[3]**
Quasi-experiments	1. Lopato, 1963	1. McConnell, 1980 2. Padilla et al., 2013 3. Lee et al., 2015 4. Fayt, 2019 5. Fleckenstein et al., 2019 6. Barbu et al., 2019*	1. Clarkson, 1992 2. Kempert et al., 2011 (mixed results)
Cross-sectional ex post facto research	2. Rafferty, 1985 3. Saunders, 1998	7. Lee, 1989	
Longitudinal ex post facto research	4. Taylor & Lafayette, 2010	8. Leon, 2006 9. Sanders, 2010 10. Fuhriman-Cleverly, 2014 11. Rega, 2015* 12. Steele et al., 2017* 13. Serafini et al., 2020	3. Han, 2012 4. Broomes, 2013 5. Choy, 2016 6. Hartanto et al., 2018
Overall Results	**100% positive**	**77% Positive**	**83% Positive**

[1]Foreign language in the elementary schools; [2]Bilingual education programs; [3]Bilingual Proficiency

* No significant relationship between Bilingualism and MA

4.1.6. Educational Contexts:

The results of this review show that bilingualism and MA has been researched across a variety of education contexts regarding the programs or conditions investigated, age groups of participants, inclusion of the control groups and outcome measures.

a. Programs or Conditions Tested

Overall results indicate that there were mainly 3 educational programs or conditions that were tested in the research related to bilingualism and MA in typically developing primary school children.

i. Bilingual Proficiency

Of the 23 studies included in this review, 6 investigated the relationship between bilingual proficiency (BP) and MA. All these studies reported a positive link between BP and MA (Clarkson, 1992; Han, 2012; Broomes, 2013; Choy, 2016; Hartanto et al., 2018) and 1 study found mixed results (Kempert et al., 2011).

ii. FLES Programs

In this review, 4 of the studies examined the impact of learning a foreign language at elementary school (FLES) on typically developing students' MA. However, the FLES programs varied in their duration and objectives, and the foreign languages taught in the programs. All these studies found positive relationship between bilingualism and MA (Lopato, 1963; Rafferty, 1985; Saunders, 1998; Taylor & Lafayette, 2010).

iii. Bilingual Education Programs

A large proportion of the studies in the review explored the effect of bilingual education programs on primary school students' MA. These BE programs included 2-way immersion (TWI), 1-way immersion, dual language immersion (DLI), transitional bilingual education (TBE), partial immersion program (PIP), and Spanish immersion program (SIP). These BE programs differed in their structure, objectives, classroom composition, bilingual proficiency assessments, and duration.

These BE and FLES programs, and bilingual contexts involved a variety of partner languages including English, Spanish, French, Russian, Japanese, Mandarin, Cantonese, Pidgin, Kreung/Tampuen, Dutch, German, and Turkish. A brief description of the BE programs has been given in chapter 2.

b. Age Groups

The results indicate that the studies involved 3 categories of age groups or grade-levels: (a) lower primary grades 1-3; (b) upper primary grades 4-6; and (c) mixed grade levels. Of the 23 studies, 7 included lower primary school students, 5 studies were with upper primary school students and 11 studies had participants that were of mixed primary grade levels.

c. Control Groups

Of the 23 included studies, 21 included a control group of monolinguals and/or students in monolingual programs. Out of these 21 studies, researchers in 4 of the studies carefully matched the control groups to enhance the internal validity of the study (Lopato, 1963; Saunders, 1998; Taylor & Lafayette, 2010; Fuhriman-Cleverly, 2014). However, 1 of the studies statistically controlled for language (Serafini et al., 2020), whereas the other developed a baseline model that served as a control group (McConnell, 1980).

d. Outcome Measures

Majority of the studies in the present review used standardized tests for assessing the students' MA that included Iowa Test of Basic Skills, Louisiana Test of Basic Skills, Comprehensive Test of Basic Skills, and Stanford achievement Test. On the contrary, in 5 of the studies, the researchers

specially developed measures of mathematics for the research purpose (Clarkson, 1992; Kempert et al., 2011; Lee et al., 2012; Fayt, 2019; Barbu et al., 2019). All the measures of mathematics ranged from simple arithmetic sums to mixed numeracy and arithmetic skills and mathematical word problems.

4.2. Findings of the Studies Reviewed

The following section presents the summaries of the individual studies included in this systematic review. It is important to note that most of these studies investigated the relationship between bilingual education and academic achievement including reading, writing, math and English language proficiency, and other cognitive benefits such as selective auditory control and executive functions However, focussing on the main objective of the present review, results of bilingualism and bilingual programs, and potential confounders have been reported pertaining to MA only. Furthermore, effect sizes (if given) are also reported only for the mathematical ability. In addition, some of the included studies extend from kindergarten to middle and high schools, but the present review reports only findings relevant to primary school students in grades 1 to 6.

4.2.1. Narrative Summaries of the Studies

1. Lopato, (1963) investigated the effects of learning conversational French for 1 year on 3rd graders' academic performance. The total number of the participants in the study was 114, recruited from a public school in New York City and a public school in suburban metropolitan New York. The study matched the participants in the experimental and control groups for age, IQ level, grade level, and SES. In addition, the teachers at the 2 participating schools were also matched for their experience, training, and ability. To assess the achievement gains of the students in both groups, the author administered Stanford Achievement Test (elementary battery) to obtain the students' pre-test and post-test scores. T-test results demonstrated that the experimental group outperformed the control group in their mean achievement gains in average arithmetic, at both schools. Moreover, there was positive correlation between the students' general academic achievement and French learning at both campuses r=.81 and r=.39).

2. McConnell (1980) investigated the impact of an Individualized Bilingual Immersion (IBI) program on Spanish-dominant and bilingual immigrants, in a quasi-experimental longitudinal study spanning 6 years from 1974-79. The IBI was in effect a Spanish/English immersion program where L1 was Spanish and L2 was English. Since it was difficult to keep track of the control group for 6 years, the author developed a baseline comparison model as an alternative of a control group. Hence, there were 390 students in the baseline model and 630 students in the IBI program followed up from preschool to 3rd grade. The sample was chosen from Washington and Texas. The author administered math subtest of Wide Range Achievement Test Battery to assess the students' mathematical achievement. For statistical analyses, the author grouped the students according to language ability. ANOVA results

show significant main effects of the IBI program attendance and ability levels for mathematics achievement favouring the IBI program and high ability students. However, the interaction effect of the program attendance and ability levels were not significant. Additionally, the statistical analyses demonstrated that after 2 years of program attendance, the students in the IBI attained average scores in math, and by the 3rd year in IBI, they performed above average on the national norms. Moreover, the main effect of gender and its interaction effect with the program attendance were not significant for math for all attendance groups on the IBI program. Likewise, the main effect and the interaction effects of bilingual skills were both significant for mathematics achievement favouring the balanced bilinguals. The author concluded that as children in IBI progressed to upper grades, they demonstrated improvement in their mathematical ability, indicating significant effect of age.

3. Clarkson (1992) compared 6th graders for bilingualism and mathematical achievement, studying at urban schools in Papua New Guinea. In the study, 232 Pidgin-speaking bilingual students from 5 government and mission-run schools served as the experimental group whereas 69 monolingual English-speaking students from 2 international and resourceful English-only schools made the control group for this study. To assess students' mathematical ability, the researchers developed valid and reliable general mathematics test and mathematical word problems test. The authors also investigated the mediating effect of 3 potential covariates: (a) cognitive development, (b) quality of housing, and (c) father's occupation. For statistical analysis, the authors carried out ANCOVAs including the language variable (language competence) that indicated the degree of bilingualism and language proficiency in the monolinguals to investigate the 'decisive' role of bilingualism on math achievement. ANCOVA analyses demonstrated that cognitive development, quality of housing and language competence all were significant covariates. Moreover, when covariates were removed, there were no significant differences between the experimental and control groups on the 2 measures of mathematical ability i.e., general mathematics test and mathematical word problems test. In other words, there was no direct effect of bilingualism on mathematical ability. In addition, Scheffe's Test results showed that the weak bilinguals performed significantly low on both measures of mathematics, suggesting the degree of bilingual proficiency a vital mediator between bilingualism and mathematical achievement. Furthermore, the significance of 'quality of housing' variable implicitly indicated the mediating role of SES.

4. Kempert et al., (2011) compared 3rd grade Turkish-German bilingual students' mathematical achievement to German monolinguals. Experimental group consisted of 44 bilinguals from 3 schools in Berlin and control group comprised of 34 students from the same 3 schools in Berlin and 2 other schools in Frankfurt. Control variables included SES, cognitive ability, and arithmetic ability. The study assessed the students' language proficiency in Turkish and German on language proficiency test (picture vocabulary/naming, synonym/antonym, verbal analogies) and a reading comprehension test only in German. For the assessment of mathematical ability, the researchers developed a math test

consisting of 2 parts: 9 mathematical word problems without distractors (WP), and 9 mathematical word problems with distractors (WP-D). Turkish students were given the math test in both German and Turkish. Monolinguals were only given German version of the math test. Kempert et al. (2011) carried out repeated measures ANOVA with the type of word problem as within-group factor (with/without distractors) and language as between group factor (monolingual/bilingual). ANOVA results show that there was no main effect of language, but there was a significant interaction effect of problem type and language (partial η^2= .06) in favour of monolinguals for solving WP. However, for solving WP-D, there was an equipoise between the two language groups (partial η^2=.00). The results of the regression analysis demonstrated that proficiency in the language of testing had a significant impact on the bilingual students' MA suggesting the advantage of attentional control for bilinguals: solving WP-D in Turkish when proficiency in Turkish was added β=.429, R^2 =.284, p<.01; and solving WP-D in German when proficiency in German was added β=.430, R^2 =.539, p<.01. Further repeated measures ANOVA with language of testing revealed a significant main effect for language of testing partial η^2=.22, a significant main effect of bilingual proficiency partial η^2=.46, and a significant interaction effect of test language and bilingual proficiency partial η^2=.42. The authors concluded that proficient bilinguals have a cognitive advantage over monolinguals and proficiency in the language of testing was an important mediator of math achievement.

5. Padilla et al. (2013) examined the impact of Mandarin immersion program on students' academic achievement. The sample for the study was recruited from an elementary school in Californian suburbs. There were a mix of native and non-native Mandarin-speaking students enrolled in the K-5 Mandarin immersion program. The authors followed up the immersion students from 2nd through 5th grade and compared their mathematics performance to non-immersion students in the same school, on state-wide California Standardized Testing and Reporting (STAR) measure. T-test results showed that in grades 2 and 3, there were no significant differences between immersion and non-immersion students in math. However, in grades 4 and 5, there were significant differences between the 2 groups in math indicating more evident cognitive advantage of bilingual education in later years at school.

6. Lee et al. (2015) compared the mathematics performance of two groups of students in rural school in Cambodia. Specifically, experimental group comprised of 50 indigenous language students, whose first language was Kreung/Tampuen, enrolled in BB[1] bilingual schools following TBE model where target language was Khmer (the official language in Cambodia). On the contrary, 50 Kreung/Tampuen speaking students enrolled in monolingual (Khmer only) state schools in similar villages in Cambodia made up the control group for the study. Bilingual program structure was consistent across all the BB bilingual schools. The authors developed a battery of math assessments (mixed skills) and administered on both groups of students every year from grade 1-4 to track their mathematical achievement. Khmer literacy and oral proficiency tests were also administered every

year. The confounders included gender, ethnicity, teacher's experience, teacher's pedagogical training. T-test analyses showed that both groups were equivalent in terms of school attendance, age, ethnicity, and parents' occupation and education level. The authors carried out mixed linear modelling, MANOVA, ANOVA, and Mann-Whitney test to investigate the differences between mathematics achievement and Khmer literacy between the two groups of students. Results of all the analyses were highly positive in favour of the bilingual education program. In addition, the study reported Partial Eta square 0.768 for school type on math performance.

7. Fayt (2019) conducted a quasi-experimental study to compare 2nd graders' mathematics achievement in bilingual and monolingual education programs. The experimental group participants were 37 Dutch-speaking students studying in bilingual (Dutch/English) classrooms, and control group participants were 39 Dutch-speaking students enrolled in monolingual (Dutch only) classrooms. The author developed a math assessment consisting of 2 types of math tasks i.e., raw arithmetic sums and word problems, all presented as auditory stimuli on a computer. Students' accuracy and response time (RT) in both groups were compared with respect to the test language and school type. The results show that overall, the experimental group performed better in L1, and scored higher in raw sums as compared to the word problems. Moreover, bilingual students outperformed monolinguals and were faster in responding to the word problems as compared with raw scores which suggested faster processing of the linguistic component of the task. However, there were no significant differences between the groups on accuracy scores on solving mathematics word problems. The effect of test language on Dutch raw sums was medium (Cohen's d = .84, and r=.39), and the effect of school type on Dutch raw sums was also medium (Cohen's d= -.51, and r=.24). Lastly, the author did not find significant differences between the groups for gender, teachers' experience, and speaking a 3rd language other than Dutch and English.

8. Fleckenstein et al. (2019) compared the mathematical ability of German-speaking students enrolled in a partial immersion program (German/English) to German-speaking students in a conventional monolingual (German only) program. There were 590 students in the partial immersion program (PIP) who were followed up from 1st grade through 4th grade for their academic performance. In the PIP, math was taught in English, but the assessment was given in German language. The latent growth curve models showed that there was accelerated improvement from grades 1-3 in immersion students' math scores. Additionally, immersion students demonstrated stronger and faster gains in mathematics in comparison with the conventional students. However, no significant effects of SES and cognitive abilities were observed.

9. Barbu et al. (2019) investigated the effects of 1-year 'content and language-integrated learning' (CLIL) program on arithmetic skills and executive functioning of native French-speaking 1st graders in France. The experimental group comprised of 59 French-speaking students learning English at 7

elementary schools in Belgium whereas control group consisted of French-speaking monolingual students at 13 elementary schools in Belgium who were not learning any foreign language. Outcome measure included executive functions in addition to arithmetic skills. The researchers developed an arithmetic test comprising of 10 addition sums and 10 subtraction sums to assess the students' mathematical ability. Chi-square test results revealed no significant differences between the experimental and control groups in terms of age, receptive vocabulary, non-verbal intelligence, bilingual program structure, and involvement in leisure activities. In addition, the authors took account of several key confounders including SES, gender, general medical history, and French and English vocabulary knowledge to match the participants in the control and experimental groups. Overall, the results indicate that there were no significant differences between the groups for arithmetic skills. Specifically, Cohen's d for addition is 0.03 in favour of monolinguals but is non-significant**,** 0.00 for subtraction, and 0.03 for auditory attention, and 0.01 for cognitive flexibility. The authors interpreted that the students developed better selective auditory attentional skills in the first year of foreign CLIL program. These auditory skills helped develop further cognitive advantage. The authors concluded that improved selective auditory skills was an important mediator of cognitive advantage of bilingualism and its effect on math ability in later years.

10. Rafferty (1985) investigated the effects of learning a foreign language in elementary school (FLES) on students' academic achievement. Students received foreign language training for 30 minutes every day. The author retrieved Louisiana Basic Skills Test results for math and language for grades 3, 4 and 5, and carried out ANCOVAs for math and language test scores. Results demonstrated that the main effect of language was significant on FLES students in math only in 5th grade, but not in grades 3 and 4. Additionally, the gender effect was significant in favour of female students, and white students outperformed blacks in mathematics. However, the interaction effects of race and gender were not significant for math.

11. Lee (1989) compared the achievement scores of Chinese American Fluent English Proficient students (FEPs) to those of non-minority English-speaking students. The author retrieved 4th, 5th, and 6th graders' test scores for reading, writing and mathematics on Comprehensive Test of Basic Skills for Spring of 1989 to investigate the differences between the experimental and control groups (N=544) with regards to their academic achievement. There were 2 subgroups of Chinese American students i.e., initially classified FEPs (FEPi) and reclassified FEPs (FEPr) based on their English language proficiency assessment. The covariates included gender, SES, L1 dialects, point of origin, number of years FEP qualified, and grade level. Results of the statistical analyses demonstrated that both groups of Chinese American performed equivalent to or above their monolingual English-speaking peers in math. Moreover, the chi-square results showed that SES, point of origin and number of FEP qualification were significant for Chinese American FEPi. However, all the covariates were non-significant for the Chinese American FEPr indicating that the bilingual instruction was effective

for Chinese American students regardless of their gender, SES and other covariates. The author concluded that Chinese American students outperformed their monolingual peers on all measures of the standardized assessments at school.

12. Saunders (1998) studies the effects of content related FLES program on 3rd graders' academic achievement. The author compared the mathematics test scores of non-FLES 3rd graders for the year 1994-95 to those of FLES 3rd graders in the year 1995-96. The FLES program offered 4 language options: French, German, Spanish, and Japanese. A block design was used to rule out the interdependence of the students in the classes. Hence, instead of treating individual participants' ITBS test scores in the statistical analyses, the author took the group means as units of analysis. Pre- and post-test difference scores were analysed in matched-pairs t-test. The results revealed that there were significant differences between the groups for mathematics achievement favouring the FLES.

13. Leon (2006) compared experimental and control group students' academic performance on Texas assessment of Knowledge and Skills (TAKS) for 3 years. There were 37 students in the experimental group enrolled in Spanish Partial Immersion Programs (SPIP) that followed a 50/50 model, in 2 elementary schools in South Texas. In SPIP, math and science were taught in Spanish whereas all other subjects were taught in English. Moreover, the SPIP structure was consistent at both campuses that participated in the study. Control groups students were those enrolled in a traditional English only (EO) monolingual educational program (n=134) at the same participating campuses. The author carried out repeated measures 2 X 3 ANOVAs to investigate the effect of the educational program (SPIP vs. EO) and time (2004/2005/2006). In addition, SPIP students were also administered Spanish version of TAKS (in addition to English version) to investigate the effect of the math test language on bilingual students' performance in math. However, 20 out of 37 students took the Spanish version of TAKS (only math and science components) in 2006. Results showed that the main program effect was significant ($p < .01$), but the time effect (TAKS performance in 2004, 2005 and 2006) and the interaction effect (time x program) were not significant for math. Further, the effect sizes (Cohen's *d*) for math were in favour of SPIP: the between-group ES is .58 for 2004, .88 for 2005, and .43 for 2006. Within-group effect sizes for math performance are .42 for 2004-5, and .12 for 2005-6, indicating the largest improvement in the 2nd year of TAKS testing when the students were in grade 4 (p. 54-55). As of the test language, the effect size was 2.29 in favour of the English version of TAKS assessment implying that the students in bilingual program did better in the language of the curriculum and customary testing.

14. Sanders (2010) compared academic achievement of 846 primary school students enrolled in a 2-way bilingual immersion program and a regular monolingual education program in Texas school district. The experimental group comprised of students in bilingual immersion program which followed 90/10 model: 90% instruction was in Spanish and 10% instruction was in English in the

beginning of the program. Students in the control group were of mixed ethnicity enrolled in a regular monolingual program where the language of instruction was only English, and no native language support was provided. The author analysed grade 3, 4, and 5 assessment data of Texas Assessment of Knowledge and Skills (TAKS) of only one cohort of students who enrolled in Spanish and English 2-way bilingual immersion program in 2002-2003 academic year. Students' demographic variables, academic growth patterns, and program type were the independent variables whereas performance on TAKS was the dependent variable. The author carried out 1-way ANOVA to investigate the effect of predictors on the outcome. Results demonstrated significant differences between 2-way bilingual immersion program and regular monolingual program students in grade 3 math, favouring 2-way bilingual immersion program. However, there were no statistically significant differences between the students' performance in math in the two educational programs in grade 4 and 5. In addition, the author did not find significant differences between both groups' performance on TAKS math component for gender and ethnicity. However, SES had significant effect on math performance such that the students on paid lunch outperformed those receiving free or reduced lunch. Furthermore, the within-group variance ANOVA results showed that the native English-speaking students (NES) outperformed the native Spanish speaking students (NSS) in math in grade 5.

15. Taylor and Lafayette (2010) compared the students' academic performance studying in FLES schools offering French or Spanish as a foreign language from 3rd to 5th grade. The authors analysed the state assessment data of control and treatment group students for 3 years for their academic performance in all subjects including math. The researchers matched the students in both groups as closely as possible taking SES and school factors into account, and carried out MANOVA, ANOVA, and MANCOVA to investigate the effects of FLES. Overall results show that the FLES students outperformed their non-FLES peers, however, the effect sizes for math were near zero.

16. Han, 2012 retrieved panel data from ECLS-K:1999 (Early childhood longitudinal atudy) to investigate the impact of bilingualism on mathematics achievement by carrying out a three-level growth curve model (time as level 1, student characteristics as level 2, and school type as level 3). Specifically, the author categorized the participants (N=16,380) into 3 language groups based on their degree of bilingual proficiency: (a) mixed bilinguals, (b) non-English-dominant bilinguals, and (c) English monolinguals. The results showed positive relationship between bilingual fluency and academic achievement. Overall results revealed that despite starting with low math scores in KG, mixed bilinguals fully close the achievement gap with their native English monolingual peers by 5th grade (Han, 2012).

17. Broomes (2013) compared the academic achievement of 3rd graders to that of 6th graders on district's standardized assessments. The key covariates included in the analyses were gender, immigration status, home language, and years lived in Canada. The results of logistic regression

analyses indicated that 3rd grade proficiency in math was a significant predictor of 6th grade proficiency in math, and all the covariates were also significant. However, the likelihood ratios showed that there were small differences in the odds for the two language groups (English monolinguals and those speaking another or additional language). Likewise, there was little difference between males' and females' predicted probability of achieving 6th grade proficiency in mathematics. The author concluded that despite the obstacles due to language, immigrant students who speak a non-English language at home were more likely to attain proficiency in math in 6th grade.

18. Fuhriman-Cleverly (2014) compared the math achievement of 5th graders enrolled in a 2-way dual language immersion program (DLI) to those in a non-DLI English-only education program in rural Western US. The author carefully matched pairs of students in experimental and control groups, controlling for age, sex, SES, ethnicity, ELL status measured by K-5 State English Language Assessment (SELA), reading ability, and dominant language. There were 53 DLI students in the experimental group studying at Los Campos Dual Language Magnet School, and 53 non-DLI students in the control group studying in 12 neighbouring schools. Students' standardized math scores on State Standards Achievement Test (SSAT) were analysed employing matched-pairs t-test. Results show that there were no significant differences between the experimental and control group students on math achievement. Moreover, the ELLs performed equally well as the native students enrolled in the non-DLI program in the neighbouring schools. This authors interpreted that the English language learners (ELLs) and limited English language proficiency students (LEPs) in bilingual education fully close the achievement gap with their native and fluent English proficient (FEP) peers in math by 5th grade. The author concluded that the non-significant results of bilingual education on mathematics achievement are in fact a positive outcome (Fuhriman-Cleverly, 2014).

19. Rega (2015) compared the mathematical achievement of 40 Spanish immersion program (SIP) students to 40 non-immersion students, in 5th grade studying at an elementary school in the US. The students in SIP received instructions in Spanish from kindergarten to 3rd grade. In 3rd grade, all the SIP and non-immersion students were administered English version of MAP test for the assessment of math and reading. T-test results showed no significant differences between the two groups in math on MAP test. The effect size was 0.40 but non-significant. Hence, it can be concluded that SIP did not have detrimental effects on students' performance in MAP mathematics. Moreover, immersion students' performance in math was not only comparable to those in traditional monolingual program, but also immersion students develop bilingual proficiency by 5th grade.

20. Choy (2016) analysed ECLS-K:1999 longitudinal data in ex post facto research to expand the scope of Han (2012)'s study by stretching the time frame (K-8) and taking in more covariates including SES, English language proficiency, students' characteristics, home and community factors, and school level factors. Choy (2016) examined the effect of home language (bilingualism/non-

English home language) on students' math IRT scale scores. There were 1,019 bilinguals and 7,017 monolinguals in the study. The results of regression analyses and ordinary least squares analyses demonstrated that without adding the covariates, there were still significant differences between monolinguals and bilinguals for their performance in math in all grades. Additionally, the author concluded that math achievement gaps in bilinguals are closed earlier by 1st grade (Choy, 2016).

21. Steele et al. (2017) investigated the effects of K-5 dual language immersion (DLI)[1] program on 7 cohorts of students in Portland public schools. There were 752 students in the experimental group enrolled in the DLI through a lottery system whereas 873 students enrolled in the traditional educational program served as the control group for this study. The intent-to-treat analysis and generalized least squares regression models demonstrated that there was positive effect of the DLI on math achievement of 5th graders, as measured on state-wide assessments. The results, though non-significant, reveal that DLI had no detrimental effects on students' mathematics achievement.

22. Hartanto et al. (2018) analysed ECLS-K:2011 cohort data to study the effects of bilingualism on 1st graders' mathematical achievement on a standardized math test. There were 2,106 bilinguals and 11,333 English-speaking monolinguals in the study. Ordinary least squares regression analyses showed that without controlling for the covariates, bilingualism had negative effect on math achievement. However, when the covariates were statistically controlled (model 2), bilingualism had significantly positive results.

23. Serafini et al (2020) studied the effects of bilingual education on academic achievement of 5 cohorts of students enrolled in two-way immersion (TWI) program in Miami School Readiness Project[1]. The students were low-income Spanish-speaking dual language learners with mixed ethnicities (N=20,870). The authors used longitudinal data on students' GPA and standardized math test scores on Florida Comprehensive Assessment Test (FCAT), from 2002 to 2007 to investigate the effects of key covariates including cognitive ability, SES, ethnicity, and socio-emotional and behavioral skills. The results of correlation analyses and regression models revealed that all covariates significantly correlated with 5th graders' math achievement, but the main effect of the TWI was very small and non-significant. However, when the authors entered the unique variable 'the speed of attaining English language proficiency', the effect of the TWI was larger and significant. Additionally, girls, high SES students, white/Asian/other ethnicities, and those with higher cognitive skills and enhanced socio-emotional and behavioral skills attained English language proficiency faster, which mediated the link between bilingual education and math achievement (Serafini et al., 2020).

4.2.2. Effect Sizes

Effect size refers to "an objective and standardized measure of the magnitude of an observed effect" which goes beyond the *p*-value and tells whether the effect is important and meaningful (Field, 2018, p. 174-5).

Of the 23 studies included in this review, only 7 report a metric of effect size (ES) explicitly. Of the 7 studies that report ES, 4 have calculated Cohen's d which is interpreted as trivial (d<.2), small (d=.2), medium (d=.5), and large (d=.8) (Cohen, 1988). The remaining 3 studies have calculated Eta square which is deciphered as small (η^2=.01), medium (η^2=.06), and large (η^2=.14).

1. Kempert et al. (2011) found no main effect for the language group but a significant interaction effect of the type of math problems and the language group (monolinguals or bilinguals) in repeated measures ANOVA and found a medium ES (partial η^2=.06) in favour of the monolinguals indicating that monolinguals outperformed bilinguals in solving math problems without distractors (WP). On the contrary, Kempert et al. (2011) found zero ES for the language group for solving problems with distractors (WP-D) showing an equipoise between the two language groups in that case. The regression analyses show that for solving WP-D in Turkish when proficiency in Turkish was added R^2 =.284, and for WP-D in German when proficiency in German was added R^2 =.539. Further ANOVA results indicate a large ES of test of language partial η^2=.22. These results highlight a positive impact of the bilingual students' proficiency in the language of math tests. (Kemper et al., 2011).
2. Lee et al. (2015)'s study reports a large ES η^2= .768 for the school type (bilingual vs. monolingual) suggesting that students in bilingual education achieved higher than those in monolingual schools.
3. Fayt (2019) found a medium[1] effect of school type on solving raw sums in Dutch d= .51, and r=.24 showing that the students in bilingual classrooms surpass those in monolingual classrooms in MA. Moreover, the effect size of math test language was also medium d=.84, and r=.39 indicating that bilingual students do better in their native language which was also the language of math instruction.
4. Barbu et al. (2019) reported that Cohen's *d* for the language group (bilingual vs. monolingual) was trivial and non-significant d=.03 favouring monolinguals for solving addition sums. However, for solving subtraction sums, the researchers reported zero ES d=.00, and trivial ES d=.03 for auditory attention, and d=.01 for cognitive flexibility in favour of bilinguals.
5. Leon (2006) found that there was a significant main effect of the bilingual program on students' MA. specifically, the author found small to large between-groups ES for SPIP[1] for the 3 years: d=.58 for 2004, d=.88 for 2005, and d=.43 for 2006. Within-group effect sizes for math performance were d=.42 for 2004-5, and d=.12 for 2005-6, meaning the largest

improvement when the students were in grade 4. In addition, the author reported large ES for the test language d=2.29 suggesting that bilingual students do better when the language of the curriculum and customary testing (Leon, 2006).

6. Taylor and Lafayette (2010) found a near zero ES for FLES program on the students' MA η^2=.00 suggesting an equipoise between the monolinguals and bilinguals in term of MA.
7. Rega (2015) reported a non-significant ES (d=.40, t=1.63, p >.1) which means that the immersion program had no detrimental effects on the students' MA and their performance in math was comparable to their monolingual peers.

4.3. Risk of Bias assessment

This systematic review has used the Mixed Methods Assessment Tool (MMAT) Version 2018 for the Risk of Bias assessment (RoBA) of each individual study. MMAT was developed by Pluye et al., (2018) for the assessment of qualitative, quantitative randomized controlled trials, quantitative non-randomized, quantitative descriptive, and mixed methods research designs. In the present review, there are quasi-experimental studies and ex post facto designs. Therefore, the 3rd component of the MMAT (2018) "Quantitative non-randomized" has been used for the RoBA of the individual studies according to the MMAT (2018) user guide.

Table 4.1 illustrates the RoBA ratings of the individual studies on the indicators of rigour given by the section 3 of MMAT (2018). For each relevant question on the MMAT (2018), the indicators of RoBA have been marked 'Yes' if the study meets the criteria/indicators and has low bias, 'No' if the study does not meet the criteria and has high bias, and 'Can't tell' if there is insufficient information related to the specific indicator. The overall weight of the evidence of each study has been ranked as 'Strong', 'Moderate' or 'Weak' in the last right column. A summary of the RoBA in table 4.1 shows that more than half of the studies give 'Moderate' evidence whereas one-third of them provide 'Strong' evidence. A detailed explanation of each indicator of bias is given below.

Table 4.3: Risk of Bias Assessment of Individual Studies

Study Reference	Quantitative non-randomized					
	Participants representative of the target population	Appropriate Intervention and outcome measures	Complete outcome data	Confounders accounted for in the design and analysis	Intervention/program administration as intended	**Global weight of the evidence rating**
	Quasi-Experiments					
1. Lopato, 1967	Yes	Yes	Yes	Yes	Yes	Strong
2. McConnell, 1980	Yes	Yes	Yes	Yes	Yes	Moderate
3. Clarkson, 1992	Yes	Yes	Yes	Yes	N/A	Moderate
4. Kempert et al., 2011	No	Yes	Yes	Yes	N/A	Moderate
5. Padilla et al., 2013	No	Yes	Yes	Yes	Yes	Moderate
6. Lee et al., 2015	Yes	Yes	Yes	Yes	Can't tell	Moderate
7. Fayt, 2019	Yes	Yes	Yes	Yes	Can't tell	Moderate
8. Fleckenstein et al., 2019	Yes	Yes	No	Yes	Can't tell	Moderate
9. Barbu et al., 2019	Yes	Yes	Yes	Yes	Yes	Moderate
	Cross-sectional Ex Post Facto Research					
10. Rafferty, 1986	Yes	Yes	Yes	Yes	Can't tell	Moderate
11. Lee, 1989	Yes	Yes	Yes	Yes	Can't tell	Moderate
12. Saunders, 1998	Yes	Yes	Yes	No	Yes	Moderate
	Longitudinal Ex Post Facto Research					
13. Leon, 2006	Yes	Yes	No	Yes	Yes	Moderate
14. Sanders, 2010	Yes	Yes	Yes	Yes	Yes	Moderate
15. Taylor & Lafayette, 2010	Yes	Yes	Yes	Yes	Can't tell	Moderate
16. Han, 2012	Yes	Yes	Yes	Yes	N/A	Moderate
17. Broomes, 2013	Yes	Yes	Yes	Yes	N/A	Moderate
18. Fuhriman-Cleverly, 2014	Yes	Yes	Yes	Yes	Can't tell	Moderate
19. Rega, 2015	Yes	Yes	Yes	Yes	Yes	Moderate
20. Choy, 2016	No	Yes	Yes	Yes	N/A	Moderate
21. Steele et al., 2017	No	Yes	Yes	Yes	Can't tell	Moderate
22. Hartanto et al., 2018	Yes	Yes	Yes	Yes	N/A	Moderate
23. Serafini et al., 2020	Yes	Yes	Yes	Yes	Can't tell	Moderate
Yes	19	23	21	22	14	**Strong: 1**
Can't tell	0	0	0	0	9	**Moderate: 22**
No	4	0	2	1	0	**Weak: 0**
Total	23					23

i. Representative Sample

With regards to the representativeness of the sample, 4 studies[4,5,20,21] demonstrated high bias due to the inappropriate control group. For instance, Kempert et al. (2011) compared Native Germans with Turkish-German immigrants. The German proficiency level of both the groups could not be equivalent. Moreover, there must be various differences between the groups (e.g., SES, immigrant status, factors related to the child and their families to name a few) which may have obscured the results, as this is the only study that has reported 'mixed' results, however interpreted as positive. Likewise, the bilingual immersion programs tested in some studies were lottery-based which implies that the students were self-motivated to learn an additional language, and the control groups were not truly monolinguals, though there was no evidence of students receiving L1 support in the non-immersion classrooms that served as the control groups (Padilla et al., 2013, Steele et al., 2017). Choy (2016) selected 40% sample selected from the nationally representative ECLS-K dataset which may have skewed the results or caused bias as the study had expanded the time frame from K-8. Therefore, only students whose data was available were included.

ii. Appropriate Intervention and Outcome Measures

Regarding appropriate intervention and outcomes measures, all the studies were ranked as having low bias because the outcomes of mathematics assessments and language proficiency were appropriate according to the intervention/program and/or the condition under scrutiny. Furthermore, the studies assessed a variety of math skills across the studies such as arithmetic, numeracy, math word problems, and standardized math tests comprising of age-appropriate mixed math skills.

iii. Complete Outcome Data

A total of 2 studies[8,13] demonstrated high bias in terms of 'complete outcome data' and the remaining 21 studies were ranked as having low bias. Fleckenstein et al. (2019) was a 4-year longitudinal study with 59% mean participation rate. Therefore, multiple imputations were done, and appropriate robust statistical procedures were carried out for statistical analysis. Likewise, in Leon (2006)'s study, for investigating the impact of the test language, the Spanish version of TAKS in math and science was administered on 5th graders, but only 54% (20 out of 37) students had taken this version. Hence, the findings of the study cannot be generalized.

iv. Confounders

Only 1 study[12] lacked inclusion of potential confounders in the design and analysis, and the remaining 22 studies demonstrated low bias on this indicator of rigor. Sanders (1998) did only take account of program attendance and grade-level.

v. Intervention/Program Administration

Overall, 9 studies[6,7,8,10,11,15,18,21,23] were marked as 'Can't tell' for the intervention/program administration because of the insufficient information. Of the remaining 14 studies, 6 studies[3,4,16,17,20,22] have been marked as 'N/A' as these studies did not involve any intervention or program delivered to the participants (Clarkson, 1992; Kempert et al., 2011; Han, 2012; Broomes, 2013; Choy, 2016; Hartanto, 2018).

Cumulative Confidence Across the Studies

As indicated in table 4.3, only 1 study gives 'Strong' evidence because of its appropriate methodology and low risk of bias. On the contrary, 22 of the studies provide 'Moderate' evidence due to the weaknesses and high bias in different components of the MMAT (2018). The results show that the insufficient information on intervention/program administration was the major source of high risk of bias in the included studies, and the sampling bias was the second big source of bias. Overall, the strength of the evidence in the set of the studies included in this review offers 'Moderate' evidence. A striking finding of the RoBA is that the study providing 'Strong' global evidence had matched the experimental and control groups (Lopato, 1963) which highlights the importance of an experimental design with matched-pairs in the sample that enhances internal validity of the study. Furthermore, it is noteworthy that 8 of the studies were rated as having low bias on all the indicators of risk in table 4.3, yet the global weight of evidence of these studies has been rated 'Moderate' for the following reasons:

i. Clarkson (1992): The experimental group participants were those studying in government mission-run schools with less qualified teachers (2-year diploma holders) and limited school facilities and resources, especially for math teaching. In contrast, the student in the control group attended highly resourceful international schools with computer facilities and teachers with better qualification (3-year diploma holders). This implies that the SES also greatly differed in the 2 groups. These student and school level factors may have obscured the results and hence, the superiority of the bilingual students in the experimental group cannot solely be attributed to BP.

ii. Rega (2015): The data was nested as the participants in both groups were selected from one school campus and ethnicity was the only control variable for the selection of the participants (2% Hispanics and 98% Caucasians). Hence, the results cannot be attributed to the Spanish Immersion Program. Moreover, the researcher did not assess the students' BP.

iii & iv. Broomes (2013) and Hartanto et al. (2018): The researchers did not assess L1 proficiency and contended with the parents' reports of the children's proficiency and use of the L1. This methodological flaw may have obscured the findings.

v. McConnell (1980): The study lacked a true control group and used a baseline model for establishing causality.

vi. Barbu et al. (2019): The CLIL structure in the experimental group across the 7 participating schools was not consistent, and the language of math instruction varied in the immersion group such that 29 students received math instruction in English, 13 in French, and 17 received bilingual instruction in math.

vii & viii. Leon (2006) and Sanders (2010): There is no evidence in the study whether the students' BP was assessed by the researchers.

4.4. Overall Results

This section presents the overall results of the review in the thematic synthesis and answers to the review questions.

4.4.1. Narrative Synthesis

i. Bilingual Proficiency and MA

Table 4.4 indicates that 6 of the 23 studies investigated the impact of bilingual proficiency on TDPSS's mathematical attainment. Out of these 6 studies, 5 (83%) found positive link between bilingualism and MA and only 1 (17%) reported mixed results. Of the studies with BP, 2 were quasi-experiments (Clarkson, 1992; Kempert et al., 2011) and 4 were causal-comparative cohort studies that involved large samples. It is noteworthy that the students in the cohort studies were not assessed for their L1 proficiency by the researchers, rather parents' reports on children's proficiency and use of L1 were considered. This aspect of the studies underscores a major methodological flaw which undermines the strength of the evidence given by these studies (table 4.4). Moreover, only 1 study in this subset has reported effect sizes (Kempert et al., 2011).

Table 4.4. Studies Investigating the Link between Bilingual Proficiency and MA

Study Reference	Age group	Student Population	Sample Size	L1	L2	Assessment of BP	Results	Effect size
Clarkson, 1992	G6	Pidgin Speaking	301	Pidgin	English	Yes	Positive	NG
Kemper et al., 2011	G3	Turkish-German vs native-German	78	Turkish	German	Yes	Mixed	η^2=.03 η^2=.00
Han, 2012	K-5	ECLS-K:1999	16,380	Mixed	English	No*	Positive	NG
Broomes, 2013	G3, G6	Immigrants in Canada	121,037	Mixed	English	No*	Positive	NG
Choy, 2016	K-8	ECLS-K:1999	8,036	Mixed	English	No*	Positive	NG
Hartanto et al., 2018	G1	ECLS-K:2011	13,439	Mixed	English	No*	Positive	NG

*Researchers contended with parents' reports to assess proficiency in L1.

ii. FLES Programs and MA

As illustrated in table 4.5, 4 of the studies examined the impact of learning a foreign language in elementary school on TDPSS' performance in mathematics, and 100% of these studies found a positive link between learning a foreign language and MA. However, only 1 study has reported effect size (Taylor & Lafayette, 2010). It is vital to note that 1 study from this subset was a quasi-experiment (Lopato, 1963) that assessed the students' BP. The remaining 3 studies were causal-comparative research and did not assess the students' BP.

Table 4.5. Studies Testing the Impact of FLES Programs on MA

Study Reference	**Age group**	**Student Population**	**Sample Size**	**L1**	**L2**	**Assessment of BP**	**Results**	**Effect size**
Lopato, 1963	G3	English speaking	114	English	French	Yes	Positive	NG
Rafferty, 1986	G3-5	English speaking	13,200	English	French	No	Positive	NG
Saunders, 1998	G3	English speaking	822	English	French, German, Spanish, Japanese	No	Positive	NG
Taylor & Lafayette, 2010	G3-5	English Speaking	4,345	English	French, Spanish	No	Positive	η^2=.00

iii. BE Programs and MA

Table 4.6 shows 13 of the studies that examined the effect of bilingual education programs on the students' MA, and 77% of these studies reported positive results. Moreover, 5 of the studies in this subset reported effect sizes of varying magnitudes. Furthermore, a variety of BE programs were tested in these studies that varied greatly with regards to the program structure, objectives, duration, and partner languages. In addition, 2 of the studies were with lottery-based BE programs and 2 were parent-choice/optional which implies that the students were self-motivated to learn the partner language. However, only in 3 of the studies, the researchers assessed the students' bilingual proficiency.

Table 4.6. Studies Examining the Relationship between BE Programs and MA

Study Reference	**Age group**	**Student Population**	**Sample Size**	**BE program**	**Partner languages**	**Assessment of BP**	**Results**	**Effect size**
McConnell, 1980	preK-3	Spanish-dominant & bilingual immigrants	1,020	IBI[1]	English	Yes	Positive	NG
Padilla et al., 2013	G2-5	Mixed languages/ ethnicities	1,029	BIP[2]*	Mandarin/ English	Yes	Positive	NG
Lee et al., 2015	G1-4	Ethnic minority	100	TBE[3]	Khmer (Kreung/ Tampuen)	No	Positive	η^2= .768
Fayt, 2019	G2	Native-Dutch	76	BI[4]	English	No	Positive	*d*= .51 *d*=.84
Fleckenstein, 2019	G1-4	Native-German	590	PIP[5]**	English	No	Positive	NG
Barbu et al., 2019	G1	Native-French	116	Foreign CLIL[6]	English	Yes	Not sig	*d*=.03 *d*=.00
Lee, 1989	G4-6	Chinese American vs English-dominant	544	BI[4]	English	No	Positive	NG
Leon, 2006	G3-5	Mixed ethnicities	171	SPIP[7]	Spanish/ English	No	Positive	Moderate to large
Sanders, 2010	G3-5	Mixed	846	BIP[2]	Spanish/ English	No	Positive	NG
Fuhriman-Cleverly, 2014	G5	Mixed minority students	106	DLI[8]	Spanish/ English	No	Positive	NG
Rega, 2015	G5	Mixed	80	SIP[9]**	Spanish/ English	No	Not sig	*d*=.40 Non-sig
Steele et al., 2017	K-8	Mixed	1,625	DLI[8*]	Spanish/ English	No	Not sig	NG
Serafini et al., 2020	G5	Spanish-speaking	20,870	TBE[3]	Spanish/ English	No	Positive	NG

[1]Individualized bilingual immersion; [2]Bilingual immersion program; [3]Transitional bilingual education; [4]Bilingual instruction; [5]Partial immersion program; [6]Content language integrated learning; [7]Spanish partial immersion program; [8]Dual language immersion; [9]Spanish immersion program

*Lottery-based immersion program; **Parent's choice

4.4.2. Review Questions

Following section presents answers to the question that guided this review.

a. Primary Question:

Q1. Do the typically developing bilingual children in primary schools have higher mathematical ability as compared with the monolinguals?

The review successfully answers the primary question. Of the 23 included studies, findings of the 20 studies support the hypothesis that the typically developing bilingual children in primary schools outperform their monolingual peers in MA at school. The remaining 3 studies found no significant relationship between bilingualism and MA in primary school children. No study reported a detrimental effect of bilingual proficiency on students' MA.

b. Secondary Questions:

Following section answers the secondary questions based on the findings of the systematic review.

Q2: Do the typically developing primary school students attending bilingual education (BE), transitional bilingual education (TBE) or foreign language in elementary school (FLES) programs demonstrate higher mathematical ability than monolingual students in English only (EO) programs?

Of the 23 studies included in this review, 13 investigated the influence of bilingual education (partial/transitional/full DLI) programs with the participants in primary schools. Majority of the studies (77%) found positive results (McConnell, 1980; Lee, 1989; Leon, 2006; Sanders, 2010; Kempert et al., 2011; Padilla et al., 2013; Fuhriman-Cleverly, 2014; Lee et al., 2015; Fayt, 2019; Fleckenstein et al., 2019) whereas 23% found no significant effect of bilingual programs on primary school students' performance in mathematics (Rega, 2015; Steele et al., 2017; Barbu et al., 2019). Furthermore, 4 studies tested the effect of FLES on students' MA, and 100% of these studies found positive results (Lopato, 1963; Rafferty, 1986; Saunders, 1998; and Taylor & Lafayette, 2010).

Q3: At what age/grade level, there is consistent math advantage of bilingualism?

Students in upper elementary grades, mostly by 5^{th} grade, demonstrate consistent positive effects of bilingual education or bilingual proficiency. Specifically, the studies with upper elementary students, i.e., 4^{th}, 5^{th} and 6^{th} graders, found 80 to 100 percent positive results (Rafferty, 1986; Lee, 1989; Clarkson, 1992; Leon, 2006; Taylor a& Lafayette, 2010;

Han, 2012; Padilla et al., 2013; Broomes, 2013; Choy, 2016; Serafini et al., 2020). However, studies with junior elementary grades 1-3 have produced mixed results (positive or no significant relationship between bilingualism and MA). In other words, the students receiving BE since kindergarten/G1develop develop enough bilingual proficiency and/or biliteracy by grades 4, 5 and 6 that is manifested in their improved performance in mathematics.

Q4. Do age and gender of the leaners have moderating effects?

Majority of the studies included in this review found moderating effect of age for bilingual advantage on MA. As students in late childhood strengthen their bilingual proficiency, they outperform their monolingual peers and close the achievement gaps (Rafferty, 1986; Lee, 1989; Clarkson, 1992; Leon, 2006; Taylor a& Lafayette, 2010; Han, 2012; Padilla et al., 2013; Broomes, 2013; Choy, 2016; Serafini et al., 2020). Speaking of gender, majority of the studies found no significant moderating effect of gender for bilingual advantage on students' performance on math assessments (McConnell, 1980; Lee, 1989; Sanders, 2010; Lee et al., 2015; Choy, 2016; Fayt, 2019). However, a small proportion of the studies found that girls outperformed boys in MA, though the difference being very small (Rafferty, 1986; Broomes, 2013; Serafini et al., 2020) whereas Hartanto et al. (2018) reported that bilingual boys achieved higher than girls in mathematics.

Q5. Do the SES and ethnicity have mediating effects?

A small proportion of the studies investigated the mediating effect of SES between bilingualism and MA and found that high SES was associated with higher achievement in mathematics (Lee, 1989; Clarkson, 1992; Sanders, 2010; Hartanto, 2018; Serafini et al., 2020). However, Fleckenstein et al. (2019) did not find significant effect of SES on MA. It is noteworthy that the researchers have used different indicators of SES which could be a reason of inconsistent results regarding the mediating effect of SES. For example, Fleckenstein et al. (2019)'s study that found non-significant effect of SES had taken account of 'parents' qualification' as a measure of SES. On the other hand, majority of the studies reporting significant effect of SES considered 'eligibility to free/reduced meals' as an indicator of SES (Lee, 1989; Sanders, 2010; Hartanto, 2018; Serafini et al., 2020) except for Clarkson (1992) that took account of 'quality of housing' as a measure of SES.

Out of the 7 studies that examined the effect of ethnicity, 5 studies found a significant effect indicating that White/Asians outperformed Blacks and Hispanics (Rafferty, 1986; Choy, 2016; Hartanto et al., 2018; Serafini et al., 2020). Specifically, Lee (1989) found that Chinese American bilinguals exceeded their English-speaking peers in MA. However, Sanders (2010) and Lee et al. (2015) reported non-significant effect of ethnicity.

Q6. Do bilingual assessments (dual language tests or test items) affect children's performance and achievement in math?

In this review, 3 of the studies investigated the impact of 'language of the math test', a vital covariate, on bilingual students' MA. All these studies reported significant effect of test language indicating that the students performed better in the language they were proficient in. Leon (2006) and Fayt (2019) found large effect sizes for math test language d=2.29 and d=.84 respectively, and Kempert et al. (2011) reported a small effect size (partial Eta square=.06).

Q7. Why are there inconsistent finding in the studies with bilingualism and MA?

Of the 23 studies included in this review, 19 found positive link between bilingualism and MA whereas 3 studies found no significant relationship and 1 study reported mixed findings. These inconsistencies in the results can be attributed to various methodological factors.

First, the age of the participant is an important factor in this regard. Prior research has shown that it takes 3-5 years to demonstrate the cognitive benefits of bilingual proficiency in students' academic achievement (Cummins, 1979). This implies that lower primary students are generally not expected to demonstrate stronger mathematics skills as students may not have developed sufficient bilingual proficiency. Therefore, studies with junior elementary students show no significant benefit of bilingualism in math (Barbu et al., 2019). However, age group solely does not define the results of bilingualism.

Second, the structure of the bilingual program is another key methodological factor in this area of research. Specifically, the percentage of L1 support, language of mathematics instruction (L1/L2), language of math assessments (L1/L2), and duration of the bilingual program and its objectives are not consistent across various research contexts, which have led to inconsistent findings. For example, Steele et al. (2017) found no significant impact of bilingual education on students' MA. It is noteworthy that in Steele et al. (2017)'s study, there were 2 models of DLI (1-way immersion (50/50) model, and 2-way immersion (90/10) model) that the immersion students were exposed to, and 4 partner languages (Russian, Japanese, Spanish, and Mandarin) in the DLI that the participants were taught on choice. These languages differ greatly in terms of language structure and orthography. Hence, the inconsistencies in the bilingual program structure make the groups of the students non-comparable which may have led to the results that are in contrast with the previous research.

Third, inclusion of potential confounders and an appropriate control group are also important methodological factors with BE research such that the studies which include an

equivalent or matched control group and take account of the covariates evidenced by prior research have reported consistently positive results. For instance, Lee (1989)'s study with Chinese American bilinguals found that SES, point of origin, and number of years since the child had attained English proficiency were significant covariates, and reported positive results of bilingualism. Similarly, Rafferty (1986) reported positive results of FLES program and found ethnicity, gender, and grade-level/age significant mediators. On the contrary, studies that don't include the confounders in the statistical analyses have reported non-significant relationship between bilingualism and MA (Rega, 2015; Steele et al., 2017).

Fourth, the assessment methods of bilingual proficiency and MA are another key factor that determine the rigor in research with bilingualism and MA. For instance, Clarkson (1992) developed valid and reliable measures of assessment of bilingual students' proficiency in the two languages (Pidgin and English), and MA (general mathematics test and mathematical word problems test) and found positive and meaningful results.

Fifth, assessment of bilingual proficiency is another factor that may have obscured the findings. For instance, some researchers assess the proficiency in the 2 languages using valid and reliable measures of vocabulary, literacy, reading or mixed language skills. Yet, some researchers contend with the parental reports of the child's use of L1 or L2. This is a major methodological inconsistency seen in this review.

Lastly, it is noteworthy that the non-significant results in the research investigating bilingualism and MA are in fact positive because non-significant results point to absence of detrimental effects of BE programs and/or bilingualism on primary school students' MA.

Conclusion

Overall results show that the studies with upper primary students demonstrated 80%-100% positive results of bilingualism for mathematical achievement. Moreover, age, SES, language background, the speed of attaining English/target language proficiency, language of mathematics instruction, and language of math test were the significant confounders of bilingual advantage in mathematical achievement. By upper primary (late childhood), the students achieved sufficient bilingual proficiency which helps them perform better on all measures of mathematical ability and closed the achievement gap with their native-English and English language proficient peers. In addition, during early years of bilingual education (grades 1-3), students began to develop attentional control and selective auditory skills that enhance their comprehension skills and receptive vocabulary. These skills mediated the link between bilingual proficiency and mathematical ability that is demonstrated in later years of primary school when students have sufficiently become biliterate. Hence, bilingualism, as Clarkson (1992) points out, is not a unidimensional phenomenon. To have cognitive

advantage of bilingualism manifested in higher mathematical ability, children must be proficient in both the languages i.e., degree of bilingual proficiency (McConnell, 1980; Clarkson, 1992; Han, 2012). Moreover, L1 support in bilingual programs assists the students attain proficiency in the target language or the language of curriculum and assessment, which in turn, enhances their comprehension of the subject matter and content knowledge in math. Therefore, the speed of attaining English language proficiency (or the target language) is a vital mediator between bilingualism and mathematical ability (Serafini et al., 2020). On the whole, bilingualism has a positive relationship with MA and does not impede typically developing primary school students' attainment in mathematics.

Chapter 5

Discussion

The review has answered all the research questions posed at the beginning of the review. The results show that bilingualism and L1 proficiency are positively linked to higher mathematical achievement when the students have attained sufficient proficiency in both L1 and L2 (balanced bilinguals) and the math test is given in the language they are proficient in. Moreover, SES and ethnicity are significant mediators between bilingualism and MA whereas age (but not gender) has a moderating effect. These results are in line with the findings of the previous research (Thomas & Collier, 2017). In addition, the research in bilingualism suffers from some fundamentals methodological challenges and flaws which yield contradictory findings. Furthermore, one reason for the dearth of research in bilingualism outside the US and colonial languages (English, French, Spanish) context is the unavailability of standardized assessments in other non-colonial languages such as Arabic, Turkish, Hindi, and Urdu, and other ethnic minority languages. A striking finding of the review is that more than two-third of the studies did not report a measure of effect size which signals a major deficit in the research.

Conclusion and Implications

A total of 23 studies, spanning over 6 decades with more than 200,000 primary school students and a total of 13 languages including English, investigating the link between bilingualism and mathematical ability, were included in this systematic review, following moderately stringent eligibility criteria described in the methodology section. The review has attempted to close the gap in research by synthesizing evidence of the prior research on bilingualism and mathematical achievement. Although, the body of research reviewed in this dissertation is small, it gives moderately strong evidence that bilingualism is positively associated with the primary school students' mathematical ability. Since most of the studies were conducted in the US, and a few outside the US, the findings of this review should be generalized with caution. Moreover, a large proportion of the minority language students in the US is Spanish-speaking, which makes the findings of the studies conducted in the US less generalizable to other parts of the world including Canada where most ELLs are French-speaking, the UK which is largely heterogeneous with respect to the students' L1, and the GCC countries where the population is mostly high SES and Arabic-speaking. The studies with ethnic minority groups in Papua New Guinea and Cambodia show promising results for preserving the students' indigenous languages alongside teaching them a foreign language with no detrimental effect on their performance in mathematics (Clarkson, 1992; Lee et al., 2015). In addition, the linguistic landscapes in African and Asian countries differ greatly, due

to the large variety of heritage languages spoken by the local population and the underdeveloped education systems and infrastructures in these countries. This situation highlights the need for conducting research in bilingualism and MA in the UK, one of the largest immigrants' destination, and the developing countries with limited resources and funding to meet the educational needs of the students.

A vital finding of this review is that bilingualism is not a simple peripheral phenomenon that it would be easy to answer the question "What is the relationship between bilingualism and mathematical ability?". Rather, it is a multidimensional complex phenomenon, and hence the answer to the above question is that it is not bilingualism that is positively, negatively or not related to mathematical ability, rather it is the 'degree of bilingual proficiency' that relates to the MA such that the 'balanced bilinguals achieve higher in math whereas 'weak bilinguals' achieve low in math. Moreover, bilingualism does not cause 'mental confusion' or 'cognitive inferiority' in ethnic minority students in schools, but the 'language barrier of L2' hinders their achievement. When the bilingual students have developed sufficient proficiency in the two languages, they begin to thrive and perform above average on the national norms. Hence, it can be concluded that the bilingual education programs that provide L1 support to the minority students are better pedagogy than submersion or English-only programs.

Another major finding of this review is that most of the studies were causal-comparative research that retrieved past assessment data to make the comparisons between the monolinguals and bilinguals to investigate the nature and strength of the bilingual advantage in the form of mathematic achievement. Not a single study reviewed here involved random assignment of the participants to the groups which is indeed impractical in bilingual research. The only study rated 'strong' on RoBA in this review was quasi-experiment that matched the control and experimental group for SES and cognitive abilities. This highlights the advantage of matched-pairs design that has high internal validity. An implication for future research, especially during the International Decade of Indigenous Languages 2022-2032 would be to design and conduct quasi-experiments in diverse linguistic contexts with carefully matched pairs of the participants to find meaningful and reliable evidence in the field of bilingualism and mathematical ability. To this end, there is a dire need of developing standardized math and language assessments in languages other than English (e.g., Arabic, Urdu, Persian, and Turkish, and other ethnic-minority languages).

Furthermore, the results show that a one-third of studies have found unique and meaningful confounders that mediate the link between bilingualism and mathematical attainment such as the 'degree of bilingual proficiency' (Clarkson, 1992; McConnell, 1980;

Han, 2012), ‘math test language’ (Leon, 2006; Kempert et al., 2011; Fayt, 2019), ‘point of origin’ (Lee, 1989), and the ‘speed of English language acquisition’ (Serafini et al., 2020). This indicates that there must be some other unknown vital mediators which can only be explored through further research in this area.

Limitations

Although this review has synthesized evidence of prior research investing the relationship between bilingualism and mathematical ability and has offered insight into the trends and findings of the research, it has some limitations which make the findings less generalizable. First, the studies only published in English were included which may have biased the results. Second, the subset of the studies with BE programs varies greatly in terms of program structure (% allocation of L1 and L2), target population, lottery-based system/optional element, aims of the program, language of math instruction, language of math test, classroom composition (linguistic diversity and ethnicity of the students), testing procedures (type of test, skills tested, language of test, assessors) and hence the findings are less comparable and generalizable. Moreover, the comparison groups in these studies were non-equivalent mostly in that the language of math instruction was not the same in both groups. Third, MMAT (2018) was not sufficient to critically appraise the quality and rigor of the research in BE because it does not include indicators that must be considered for the critical appraisal for research in this area such as appropriateness of the statistical analyses, language of the outcome measures, and language of instruction for the outcome variable. One suggestion is to develop or adapt the critical appraisal tool for research with bilingualism by including more categories/indicators of rigor to assess the substantial bias.

References

Barbu, C., Gonzalez, A., Gillet, S., & Poncelet, M. (2019). Cognitive advantage in children enrolled in a second-language Immersion Elementary School Program for one year. *Psychologica Belgica*, *59*(1), 416–435. https://doi.org/10.5334/pb.469

Bialystok, E. & Hakuta, K. (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. Birdsong, D (Ed.), Second language acquisition and critical period hypothesis (pp. 161-181). Lawrence Earlbaum associates.

Broomes, O. P. (2013). More than a new country: Effects of immigration, home language, and school mobility on elementary students' academic achievement over time. Education Policy Analysis Archives, 21(48).

Choy, y. Y. (2016). *Does bilingualism improve academic performance? Estimating the relationship between foreign language spoken at home and student test scores* (Master's thesis). Available from ProQuest Dissertations and Theses database.

Clarkson, P. C., & Galbraith, P. (1992). Bilingualism and mathematics learning: Another perspective. *Journal for Research in Mathematics Education. 23(1)*, 34-44.

Collier, V. P., & Thomas, W. P. (2017). Validating the power of bilingual schooling: Thirty-two years of large-scale, Longitudinal Research. *Annual Review of Applied Linguistics*, *37*, 203–217. https://doi.org/10.1017/s0267190517000034

Critical period hypothesis. Psynso. (n.d.). Retrieved August 10, 2022, from https://psynso.com/critical-period-hypothesis/

Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. Working Papers on Bilingualism, 9, l-43.

Cummins, J. (1981). Age on arrival and Immigrant Second Language Learning in Canada: A reassessment. *Applied Linguistics*, *2*(2), 132–149. Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review*, *56*(1), 18–37. https://doi.org/10.17763/haer.56.1.b327234461607787

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review*, *56*(1), 18–37. https://doi.org/10.17763/haer.56.1.b327234461607787

Cummins, J. (1992). Bilingual education and English immersion: The ramírez report in theoretical perspective. *Bilingual Research Journal*, *16*(1-2), 91–104. https://doi.org/10.1080/15235882.1992.10162630

Databases. EPPI. (2022, August 11). Retrieved August 11, 2022, from https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=185

Data Extraction Forms. Data extraction forms | Cochrane Developmental, Psychosocial and Learning Problems. (n.d.). Retrieved August 12, 2022, from https://dplp.cochrane.org/data-extraction-forms

Education Endowment Foundation. EEF. (n.d.). Retrieved August 11, 2022, from https://educationendowmentfoundation.org.uk/

Encyclopedia of Research Design. SAGE Research Methods. (2012, December 27). Retrieved August 12, 2022, from https://methods.sagepub.com/reference/encyc-of-research-design/n353.xml

Evans, J., & Benefield, P. (2001). Systematic reviews of educational research: Does the medical model fit it?. *British Educational Research Journal*, *27*(5).

Fleckenstein, J., Gebauer, S. K., & Möller, J. (2019). Promoting mathematics achievement in one-Way Immersion: Performance Development over four years of elementary school. *Contemporary Educational Psychology*, *56*, 228–235. https://doi.org/10.1016/j.cedpsych.2019.01.010

Fayt, C. (2019). *Achievement on arithmetic in bilingual primary schools* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UNI No. 3888789)

Fuhriman-Cleverly, V. (2014). *The academic and linguistic effectiveness of a dual language program: A comparison of student achievement within one district* (Doctoral thesis). Available from ProQuest Dissertations and Thesis database. (UMI No. 3643020)

Greene, J. (1998). A Meta-Analysis of the Effectiveness of Bilingual Education.

Han, W.-J. (2011). Bilingualism and academic achievement. *Child Development*, *83*(1), 300–321. https://doi.org/10.1111/j.1467-8624.2011.01686.x

Hakuta, K. (1986). *Mirror of language: The debate on bilingualism*. Basic books, Inc.

Hartanto, A., Yang, H., & Yang., S. (2018). Bilingualism positively predicts mathematical competence: Evidence from two large-scale studies. *Learning and individual differences*. *22(3)*, 269-428.

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from handbook.cochrane.org

Hong, Q. N., & Pluye, P. (2018). A conceptual framework for critical appraisal in systematic mixed studies reviews. *Journal of Mixed Methods Research*, *13*(4), 446–460. https://doi.org/10.1177/1558689818770058

Intelligent Systematic Review. Rayyan. (n.d.). Retrieved August 11, 2022, from https://www.rayyan.ai/

International Database of Education Systematic Reviews. IDESR. (n.d.). Retrieved August 11, 2022, from https://idesr.org/

Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology*, *103*(3), 547–561. https://doi.org/10.1037/a0023619

Lee, E. W. (1989). Chinese American fluent English proficient students and school achievement. *NABE Journal*, *13*(2), 95–111. https://doi.org/10.1080/08855072.1989.10668553

Lee, S., Watt, R., & Frawley, J. (2014). Effectiveness of bilingual education in Cambodia: A longitudinal comparative case study of ethnic minority children in bilingual and monolingual schools. *Compare: A Journal of Comparative and International Education*, *45*(4), 526–544. https://doi.org/10.1080/03057925.2014.909717

Leon, L. D., (2006). *The effectiveness of a Spanish partial immersion program among fifth graders in South Texas* (Doctoral thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3239707)

Lopato, E. W., (1963). FLES and academic achievement. *The French Review*, *36*(5), 499-507.

McConnell, B. B., (1980). *Effectiveness of Individualized bilingual instruction for immigrant students* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 48106)

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1). https://doi.org/10.1186/2046-4053-4-1

Newman, M., & Gough, D. (2019). Systematic Reviews in Educational Research: Methodology, Perspectives and Application. *Systematic Reviews in Educational Research*.

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/english two-Way Immersion Program: Language proficiency and academic achievement. *Foreign Language Annals*, *46*(4), 661–679. https://doi.org/10.1111/flan.12060

Prisma. PRISMA. (n.d.). Retrieved August 11, 2022, from https://prisma-statement.org/

Prisma. (n.d.). Retrieved August 11, 2022, from https://prisma-statement.org/PRISMAStatement/FlowDiagram

Prospero. (n.d.). Retrieved August 11, 2022, from https://www.crd.york.ac.uk/PROSPERO/searchquick.php

Rafferty, E. A., (1986). Second language study and basic skills in Louisiana.

Rega, M. K. (2015). *Reading and math outcomes of randomly selected majority culture students participating in an elective, parent choice, full academic content area Spanish immersion program* (Doctoral thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3687022)

Rossell, C. H. & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English*, *30*(1), 7-74.

Saunders, C. M., (1998). *The effect of the study of a foreign language in the elementary school on scores on the Iowa Test of Basic Skills and an analysis of student-participant attitudes and abilities* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 9836979)

Sanders, A. N. (2010). *The effectiveness of two-way bilingual immersion programs in closing the achievement gap for minority students* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3447332)

Search & Browse. (n.d.). Retrieved August 11, 2022, from https://www.cochrane.org/search/site

Serafini, E. J., Rozell, N., & Winsler, A. (2020). Academic and English language outcomes for dlls as a function of school bilingual education model: The role of two-way immersion and home language support. *International Journal of Bilingual Education and Bilingualism*, *25*(2), 552–570. https://doi.org/10.1080/13670050.2019.1707477

Skutnabb-kangas, T. (2012). Mother-tongue-medium education. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0776

Skutnabb-Kangas, T. (2012). Cummins, Jim. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0306

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on Student Achievement. *American Educational Research Journal*, *54*(1_suppl). https://doi.org/10.3102/0002831216634463

Taylor, C., & Lafayette, R. (2010). Academic achievement through FLES: A case for promoting greater access to foreign language study among young learners. *The Modern Languages Journal. 94(i)*, 22-42.

The 10 largest refugee crises to follow in 2022. Concern Worldwide. (2022, April 25). Retrieved August 11, 2022, from https://www.concernusa.org/story/largest-refugee-crises/

Tsimpli, I., Mukhopadhyay, L., Treffers-Daller, J., Alladi, S., Marinis, T., Panda, M., Balasubramanian, A., & Sinha, P. (2019). Multilingualism and multiliteracy in primary education in India: A discussion of some methodological challenges of an interdisciplinary research project. *Research in Comparative and International Education*, *14*(1), 54–76. https://doi.org/10.1177/1745499919828908

World migration report 2022: Selected infographics. World Migration Report. (n.d.). Retrieved August 11, 2022, from https://worldmigrationreport.iom.int/world-migration-report-2022-selected-infographics

Appendix A: IDESR Registration Form

Title:

Bilingualism and mathematical Ability in Typically Developing Primary School Children: A Systematic Review

Review Questions:

Q1. Do the typically developing bilingual children in primary schools have higher mathematical ability as compared with the monolinguals?

Q2. Do the typically developing primary school children attending bilingual education programs (BE), transitional bilingual programs (TBE), and foreign language in the elementary school (FLES) programs demonstrate higher mathematical achievement as compared to those attending English only (EO) programs in primary schools?

Q3. At what age/grade level is there consistent math advantage of bilingualism?

Q4. Do age and gender of the leaners have moderating effect?

Q5. Do the SES and ethnicity have mediating effects?

Q6. Do bilingual math assessments (dual language math tests) affect children's achievement in math?

Q7. Why are there conflicting or inconsistent findings in the studies with bilingual education and mathematics achievement?

Rationale:

A large body of research has been dedicated to studying cognitive advantages of bilingualism in all ages. However, the research investigating the link between bilingualism and mathematical attainment in young learners is limited. This systematic review aims to investigate the relationship between bilingualism and mathematical attainment in primary school children (aged 5-11) and assessing the roles of moderators (age and gender) and mediators (SES and ethnicity) of bilingualism on mathematical ability. Due to their strong economies and better opportunities for education and career, there is a trend of immigration to the English-speaking countries. According to UNO's World Migration Report published in 2020, more than half of the international migrants (141M) lived in Europe and North America, where Canada being the largest resettlement destination for the refugees and the USA the 2nd largest. This implies that the number of bilingual immigrant children is increasing in these countries which reflects the need for research to deepen our understanding of the link between bilingualism and mathematical achievement, as mathematics is one of the core subjects taught at elementary schools and mathematical attainment is vital to children's academic achievement. This review will provide insight into how bilingualism links to mathematical attainment which in turn, can potentially be useful for implementing effective pedagogy with implications of including mother tongue in the primary school classrooms comprising of students from diverse ethnic and linguistic backgrounds.

Inclusion Criteria:

1. Population: Studies with primary school children in grades 1-6

2. Outcomes measured: Studies that have language related outcomes (e.g., reading, vocabulary, spelling, grammar etc.) and mathematical ability

3. Study Designs: Longitudinal, cross-sectional, correlational, intervention, small scale and pilot studies that are empirical research and measure the language ability and mathematical attainment of primary school children.

4. Geography: Anywhere

6. Publication type: Open access Academic journals, doctoral and masters' theses and dissertation, program evaluation reports that are published, peer-reviewed and scientific pieces of empirical research, books, book chapters, and conference papers.

7. Time limit: No

8. Language in which studies were published: English

Information Sources

Following electronic bibliographic databases will be searched:

EBSCOhost

Web of Science

PsycINFO

ProQuest

ProQuest Dissertations and Theses

MEDLINE

SCOPUS

Search Strategy

The search strategy will combine terms that relate to bilingualism with terms that relate to mathematical attainment and numeracy. In addition, we will include the key search terms that relate to primary school children (e.g., elementary school, grade 1, grade 2 and so on). Search terms will be adapted for use with other bibliographic databases. There will be no date restrictions. In addition to the bibliographic database searches, we will take account of all studies included in previous meta-analyses, will conduct a manual review of the table of contents for key journals and will scan reference lists in short-listed papers.

Below is the Boolean phrase of one of the electronic bibliographic databases.

EBSCOhost search syntax

Databases searched in EBSCO:

a. British Education Index

b. Teacher Reference Centre

c. Childhood and Adolescent Studies

d. ERIC

Search String/ Boolean phrase

(bilingual* OR multilingual OR foreign language learn* OR second language learn* OR additional language learn* OR first language fluency OR mother tongue fluency) AND (math* OR numeracy) AND (attainment OR achievement OR success OR ability OR

understanding) AND (primary OR elementary OR first-grade OR second-grade OR third-grade OR fourth-grade OR fifth-grade) NOT (middle school OR high school OR secondary school OR preschool OR sixth-grade OR seventh-grade OR eighth-grade OR ninth-grade OR tenth-grade)

Search mode and expanders:

Boolean phrase

Apply related words

Also search full text of the articles

Apply equivalent subject

Limit search to

Publication date: no limit

Location in the text: All Text

Scholarly/ peer reviewed

Full text

Special limiters for British education index:

Publication type: Academic journals

Document Type: all

Language: English

Education Level: Elementary Education

Age level: Children

Special limiters for Child development and adolescent studies

Publication type: Academic journals

Document Type: all

Special limiters for Teacher Reference Centre:

Number of pages: all

Full text

Special limiters for ERIC:

Journal or document type: All

Education Level: Elementary Education

Publication type: All

Intended audience: All

Language: English

What Works Clearinghouse Reviewed: All

Search results: 46

Link to the search results:

https://ezproxy-prd.bodleian.ox.ac.uk:2713/login.aspx?direct=true&AuthType=ip,uid&db=bri&db=fgh&db=eric&db=trh&bquery=TX+(bilingual*+OR+multilingual+OR+foreign+language+learn*+OR+second+language+learn*+OR+additional+language+learn*+OR+first+language+fluency+OR+mother+tongue+fluency)+AND+(math*+OR+numeracy)+AND+(attainment+OR+achievement+OR+success+OR+ability+OR+understanding)+AND+(primary+OR+elementary+OR+first-grade+OR+second-grade+OR+third-grade+OR+fourth-grade+OR+fifth-grade)+NOT+(middle+school+OR+high+school+OR+secondary+school+OR+preschool+OR+sixth-grade+OR+seventh-grade+OR+eighth-grade+OR+ninth-grade+OR+tenth-grade)&cli0=FT&clv0=Y&cli1=RV&clv1=Y&dli0=PT100&dlv0=Academic+Journal&dld0=bri&dli1=LA10&dlv1=English&dld1=bri&dli2=DE3&dlv2=Elementary+Education&dld2=bri&dli3=AG3&dlv3=Children&dld3=bri&dli4=PT10&dlv4=academic+journal&dld4=fgh&dli5=DE3&dlv5=Elementary+Education&dld5=eric&dli6=LA99&dlv6=eng&dld6=eric&dli7=FM&dlv7=Y&dld7=trh&type=1&searchMode=Standard&site=ehost-live&authtype=ip,uid

Data Management:

Search results from the 7 databases will be exported to EndNote in RIS format. Next these references will be exported to Rayyan data management software. Abstract screening will be completed in Rayyan. Full texts of the studies shortlisted abstract screening will be downloaded and screened thoroughly. Excel software will be used to extract data from the selected studies. Statistical analyses will be carried out in SPSS 28.

Selection Process:

A. Initial Screening

In Rayyan software, after removing the duplicates, titles and abstracts retrieved using the electronic search strategy and those from additional sources will be subjected to an initial screening by the first author to determine relevance, using the following three questions:

1. Is the study a primary experimental study?

2. Does the reference include measurement of participants' mathematical ability?

3. Does the reference include an experimental group of bilinguals (English language learners or dual language learners), or refer to measures related to language ability.

If answers to any of the above three questions are “No”, the reference

will be excluded at this stage.

B. Full text screening

The first author will scan the full texts of the shortlisted studies. The following questions will be answered to evaluate the relevance of the full texts:

1. Does the study include group means and standard deviations to enable a measure of difference between groups to be calculated. Alternatively, does it include a measure of effect size (r) for the relationship between implicit learning and language in correlational studies.

2. Does it include data on at least one of the following language measures: Grammar (eg: TROG-2), vocabulary knowledge, phonological awareness, reading, written language, either used as a variable in its own right or to classify or justify a classification of participants into groups.

3. Does the study include an appropriate control group of monolinguals or the students studying in monolingual programs? Alternatively, does the study control for language in the statistical analyses?

All records will be dual screened by the first and second authors. Any inconsistencies and disagreements will be resolved by discussion and consensus. The screening process will be blinded to each screener in Rayyan software.

Data Collection Process:

Prior to data extraction, a full coding scheme with all relevant variables will be made by the first author. For this, a data extraction form will be designed and piloted on approximately 5 studies. Excel software will be used to extract data. To ensure coding is reliable, the first 25% of the studies for each implicit learning paradigm will be double coded by the first and second author. The two screeners will examine any differences in coding and resolve any disagreements with discussion and by consulting the original studies before continuing coding.

In case of insufficient information in the published study/report, the authors will be contacted directly.

Data Items:

Data items according to the following 4 headings/categories will be extracted:

1. Study Features (study name (authors); year of publication; funding sources (if any); design; participants' information (age, gender, SES, and ethnicity); sample size; sampling techniques used; subgroups within the study (including control groups); assigned language disorder if any.

2. Implicit learning task features (mathematical ability scores; modality (non-verbal or verbal tests/assessments); length; number of repetitions of sequence / stimulus)).

3. Language measures (test name; standardised or researcher-made; type of test (language- or literacy-related)

4. Statistics (standardised mean difference between groups or correlation between implicit learning and language measures).

Risk of bias/Trustworthiness of the studies

As a variety of studies are expected to be retrieved in the process of electronic search, appropriate tools of risk of bias assessment will be used to assess the trustworthiness of the studies. Cochrane Risk of Bias Tool will be used to assess the trustworthiness of the intervention studies. For non-randomized quantitative studies and longitudinal and correlational studies, Mixed Methods Appraisal Tool (MMAT; Pluye et al., 2018) will be used to assess the trustworthiness of the studies. MMAT is used for assessing the studies on qualitative, quantitative and mixed methods criteria (Pluye et al. 2018). MMAT is appropriate for critical appraisal of qualitative, RCTs, non-RCT quantitative, quantitative descriptive, and, mixed methods studies. The critical appraisal of the studies will provide information on numerical data and summaries of variables of interest, and quality of the study design. This information will be used to compose the quantitative and/or qualitative syntheses.

Data Synthesis:

Depending on the data extracted from the studies i.e., comparability of the study design, population, outcome measures and statistics, a meta-analysis will be synthesized in SPSS version 28. Cohen' d and/or Pearson's r will be used to summarize the effect size of the studies. If a quantitative analysis is not possible, a thematic synthesis will be composed.

Meta Biases

In order to address the publication bias, grey literature (dissertations and theses of postgraduate and doctoral students) will also be included.

Confidence in Cumulative Evidence:

In order to assess the strength of the body of evidence in this systematic review, GRADE assessment of the cumulative evidence will be carried out.

Specifically, following steps (according to the GRADE assessment) will be taken:

1. Take into account the limitations of the study
2. Assess the publication bias
3. Assess the consistency or inconsistency of the findings of the study
4. Precision of the estimates of the effect sizes
5. Assess the risk of bias
6. Assess the sampling bias if any
7. How the evidence accumulated answers the research questions of the review
8. Assess the intervention fidelity in the intervention studies included in this systematic review

Sources of Funding:

None

Anticipated Start Date: 20-05-2022

Anticipated End Date: 10-08-2022

Appendix B: List of Included Studies

Barbu, C., Gonzalez, A., Gillet, S., & Poncelet, M. (2019). Cognitive advantage in children enrolled in a second-language Immersion Elementary School Program for one year. *Psychologica Belgica*, *59*(1), 416–435. https://doi.org/10.5334/pb.469

Broomes, O. P. (2013). More than a new country: Effects of immigration, home language, and school mobility on elementary students' academic achievement over time. Education Policy Analysis Archives, 21(48).

Choy, y. Y. (2016). *Does bilingualism improve academic performance? Estimating the relationship between foreign language spoken at home and student test scores* (Master's thesis). Available from ProQuest Dissertations and Theses database.

Clarkson, P. C., & Galbraith, P. (1992). Bilingualism and mathematics learning: Another perspective. *Journal for Research in Mathematics Education. 23(1),* 34-44.

Fleckenstein, J., Gebauer, S. K., & Möller, J. (2019). Promoting mathematics achievement in one-Way Immersion: Performance Development over four years of elementary school. *Contemporary Educational Psychology*, *56*, 228–235. https://doi.org/10.1016/j.cedpsych.2019.01.010

Fayt, C. (2019). *Achievement on arithmetic in bilingual primary schools* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UNI No. 3888789)

Fuhriman-Cleverly, V. (2014). *The academic and linguistic effectiveness of a dual language program: A comparison of student achievement within one district* (Doctoral thesis). Available from ProQuest Dissertations and Thesis database. (UMI No. 3643020)

Han, W.-J. (2011). Bilingualism and academic achievement. *Child Development*, *83*(1), 300–321. https://doi.org/10.1111/j.1467-8624.2011.01686.x

Hartanto, A., Yang, H., & Yang., S. (2018). Bilingualism positively predicts mathematical competence: Evidence from two large-scale studies. *Learning and individual differences*. *22(3),* 269-428.

Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology*, *103*(3), 547–561. https://doi.org/10.1037/a0023619

Lee, E. W. (1989). Chinese American fluent English proficient students and school achievement. *NABE Journal*, *13*(2), 95–111. https://doi.org/10.1080/08855072.1989.10668553

Lee, S., Watt, R., & Frawley, J. (2014). Effectiveness of bilingual education in Cambodia: A longitudinal comparative case study of ethnic minority children in bilingual and monolingual schools. *Compare: A Journal of Comparative and International Education*, *45*(4), 526–544. https://doi.org/10.1080/03057925.2014.909717

Leon, L. D., (2006). *The effectiveness of a Spanish partial immersion program among fifth graders in South Texas* (Doctoral thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3239707)

Lopato, E. W., (1963). FLES and academic achievement. *The French Review*, *36*(5), 499-507.

McConnell, B. B., (1980). *Effectiveness of Individualized bilingual instruction for immigrant students* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 48106)

Mixed methods appraisal tool (MMAT) version 2018 - pbworks. (n.d.). Retrieved August 12, 2022, from http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/english two-Way Immersion Program: Language proficiency and academic achievement. *Foreign Language Annals*, *46*(4), 661–679. https://doi.org/10.1111/flan.12060

Rafferty, E. A., (1986). Second language study and basic skills in Louisiana.

Rega, M. K. (2015). *Reading and math outcomes of randomly selected majority culture students participating in an elective, parent choice, full academic content area Spanish immersion program* (Doctoral thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3687022)

Saunders, C. M., (1998). *The effect of the study of a foreign language in the elementary school on scores on the Iowa Test of Basic Skills and an analysis of student-participant attitudes and abilities* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 9836979)

Sanders, A. N. (2010). *The effectiveness of two-way bilingual immersion programs in closing the achievement gap for minority students* (Doctoral Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3447332)

Serafini, E. J., Rozell, N., & Winsler, A. (2020). Academic and English language outcomes for dlls as a function of school bilingual education model: The role of two-way immersion and home language support. *International Journal of Bilingual Education and Bilingualism*, *25*(2), 552–570. https://doi.org/10.1080/13670050.2019.1707477

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on Student Achievement. *American Educational Research Journal*, *54*(1_suppl). https://doi.org/10.3102/0002831216634463

Taylor, C., & Lafayette, R. (2010). Academic achievement through FLES: A case for promoting greater access to foreign language study among young learners. *The Modern Languages Journal. 94(i)*, 22-42.

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

Appendix C: Data Extraction Form (adapted from Cochrane Data Extraction Form for RCTs and non-RCTs, n.d.)

Review title or ID	
Study ID	

General Information

Date form completed *(dd/mm/yyyy)*	
Name/ID of person extracting data	
Reference citation	
Study author contact details	
Publication type *(e.g. full report, abstract, letter)*	

Study eligibility

Study Characteristics	Eligibility criteria *(Insert inclusion criteria for each characteristic as defined in the Protocol)*	Location in text or source *(pg & ¶/fig/table/other)*
Type of study	Randomised Controlled Trial	
	Quasi-randomised Controlled Trial	
	Other design (specify):	
Participants		
Types of intervention		
Types of comparison		
Types of outcome measures		

Participants

	Description	Location in text or source
Population description *(from which study participants are drawn)*		

Setting *(including location and social context)*		
Inclusion criteria		
Exclusion criteria		
Method of recruitment of participants *(e.g. phone, mail, clinic patients)*		
Clusters *(if applicable, no., type, no. people per cluster)*		
Baseline imbalances		
Age		
Sex		
Race/Ethnicity		
Other relevant sociodemographic		
Subgroups measure		
Subgroups reported		

Educational Program/Experimental conditions

	Description as stated in report/paper	Location in text or source
Description of the program		
Duration of treatment period		
Timing		
Administration Fidelity		

Outcome 1

	Description as stated in report/paper	Location in text or source
Type and name of the assessment measure		

Psychometric properties of the assessment		
Outcome		
Time points measured		
Person measuring/ reporting		
Unit of measurement*		
Assumed risk of bias		

*Raw scores, standardized scores, RIT scores, Grade point average, NCE scores, grade equivalent scores etc.

Outcome 2

	Description as stated in report/paper	Location in text or source
Type and name of the assessment measure		
Psychometric properties of the assessment		
Test language		
Outcome		
Time points measured		
Person measuring/ reporting		
Unit of measurement*		
Assumed risk of bias		

*Raw scores, standardized scores, RIT scores, Grade point average, NCE scores, grade equivalent scores etc.

Outcome 3

	Description as stated in report/paper	Location in text or source
Type and name of the assessment measure		
Psychometric properties of the assessment		
Test language		
Outcome		

Time points measured		
Person measuring/ reporting		
Unit of measurement*		
Assumed risk of bias		

*Raw scores, standardized scores, RIT scores, Grade point average, NCE scores, grade equivalent scores etc.

Data Analysis

	Description as stated in report/paper	Location in text
Statistical controls		
Subgroup		
Time point		
Any other results reported *(e.g. odds ratio, risk difference, CI or P value)*		
Missing data %		
Imputations		
Unit of analysis *(by individuals, cluster/groups or body parts)*		
Statistical methods used and appropriateness of these *(e.g. adjustment for correlation)*		
Results		
Effect size		
Key conclusions of study authors		