ENGLISH ACADEMIC VOCABULARY OF DUAL LANGUAGE PROGRAMME AND NON-DUAL LANGUAGE PROGRAMME SECONDARY SCHOOL PUPILS IN SELANGOR

WONG EK YING

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

ENGLISH ACADEMIC VOCABULARY OF DUAL LANGUAGE PROGRAMME
AND NON-DUAL LANGUAGE PROGRAMME SECONDARY
SCHOOL PUPILS IN SELANGOR

WONG EK YING

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF EDUCATION
(ENGLISH LANGUAGE EDUCATION)

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Wong Ek Ying**

Matric No: **PGP150014**

Name of Degree: **Master of Education (English Language Education)**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

ENGLISH ACADEMIC VOCABULARY OF DUAL LANGUAGE PROGRAMME AND NON-DUAL LANGUAGE PROGRAMME SECONDARY SCHOOL PUPILS IN SELANGOR

Field of Study: **English Language Education**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature   Date:

Subscribed and solemnly declared before,

Witness's Signature   Date:

Name:

Designation:

ABSTRACT

Content and Language Integrated Learning (CLIL) is regarded as a fertile ground for target language acquisition. However, there is a growing skepticism towards the novelty of CLIL due to various methodological flaws in the past studies. The purpose of this study is to investigate if CLIL approach truly contributes to English academic vocabulary learning when three confounding factors: selection effect, total English instruction hours and extramural English exposure are controlled. This study employs cross-sectional research design using both quantitative and qualitative data collection methods by comparing the academic vocabulary performance between secondary students in two different learning environments — regular mainstream classes and Dual Language Program (DLP), along with an interview with a DLP teacher to add depth to the overall analysis. This study involves 72 Form 2 and Form 4 students from two schools in Selangor. The results of Academic Word List (AWL) test are analyzed using independent t-tests; it is shown that there is no significant difference between DLP and non-DLP students. In other words, DLP students do not outperform non-DLP. However, it is note-worthy that the DLP group manages to stay on par with senior non-DLP students who are two years older, showing that DLP has yet to live up to its full potential. Issues hampering DLP students from reaping the full benefits of CLIL and implications of the study are therefore discussed. Although DLP does not exhibit impressive outcome in this study, the researcher is optimistic of its endeavor in Malaysia because if implemented properly, DLP is believed to have a lot more to offer, especially in academic language learning.

Keywords: CLIL, DLP, Academic Vocabulary, CALP, AWL

KOSA KATA AKADEMIK BAHASA INGGERIS ANTARA MURID KELAS PROGRAM DWIBAHASA DENGAN KELAS BIASA SEKOLAH MENENGAH DI SELANGOR

ABSTRAK

Content and Language Integrated Learning (CLIL) dianggap sebagai program yang berkesan untuk pemerolehan bahasa sasaran. Walau bagaimanapun, keberkesanan CLIL sentiasa diragui disebabkan oleh kelemahan metodologi dalam kajian lepas. Tujuan kajian ini adalah untuk menyiasat jika pendekatan CLIL menjadi penyumbang kepada pembelajaran kata akademik apabila kesan pemilihan, jumlah masa pengajaran dalam bahasa Inggeris dan pendedahan Bahasa Inggeris luar sekolah dikawal. Kajian ini dilaksanakan secara keratan rentas dengan kaedah campuran untuk pengumpulan data. Prestasi kosa kata akademik telah dibandingkan antara murid sekolah menengah dalam dua persekitaran pembelajaran, iaitu kelas biasa dan kelas Dual Language Program (DLP). Temu ramah dengan seorang guru DLP telah pun digunakan untuk analisis data yang lebih menyeluruh. Kajian ini melibatkan 72 orang murid Tingkatan 2 dan Tingkatan 4 dari dua buah sekolah di Selangor. Hasil ujian Academic Word List (AWL) daripada analisis ujian T sampel tak bersandar menunjukkan bahawa tiada terdapat perbezaan yang signifikan antara murid DLP dan bukan DLP. Namum begitu, kumpulan DLP telah berjaya mencapai keputusan yang setanding dengan murid bukan DLP yang dua tahun lebih tua. Hal ini menunjukkan bahawa DLP masih belum mencapai potensi sepenuhnya. Oleh itu, isu-isu penghalang semasa pelaksanaan DLP dan implikasi kajian telah dibincangkan. Walaupun DLP tidak menunjukkan hasil yang mengagumkan kali ini, penyelidik berasa optimis terhadap masa depan DLP di Malaysia kerana jika dilaksanakan dengan sistematik, DLP mempunyai banyak potensi terutama dalam pembelajaran bahasa akademik.

Kata kunci: CLIL, DLP, Kosa kota akademik, CALP

ACKNOWLEDGEMENT

First of all, I would like to take this opportunity to express my sincere gratitude to all my lecturers in Faculty of Education, University Malaya, especially to my supervisor, Associate Prof. Dr. Juliana Othman for all the rendered guidance, patience and assistance that enabled me to complete this dissertation smoothly.

I would like to dedicate this humble work to my husband, my parents, my parents-in-law and Bangsar Growth Group members for their relentless support, encouragement and accountability.

Most importantly, all praise and glory to God, the creator of all, who sent His Son Jesus Christ to propitiate for our sins and take the punishment sinners like us deserve. Jesus Christ is the true son of God, the promised King and to Him I owe my life.

CONTENTS **Page**

Chapter 1: Introduction

Chapter 4: Data Analysis

Chapter 5: Discussion and Conclusion

LIST OF FIGURES

-

LIST OF TABLES

LIST OF ABBREVIATIONS

AWL   Academic Word List

BICS   Basic Interpersonal Communication Skills

BPPDP   Bahagian Perancangan dan Penyelidikan Dasar Pendidikan

CALP   Cognitive academic language proficiency

CBI   Content-based Instruction

CLIL   Content and Language Integrated Learning

DLP   Dual Language Program

ESL   English as a Second Language

FL   Foreign Language

GPS   Gred Purata Sekolah

JPN   Jabatan Pelajaran Negeri

L2   Second Language

MoE   Ministry of Education

SLA   *Second Language Acquisition*

UPSR   *Ujian Pencapaian Sekolah Rendah*

VLT   Vocabulary Level Test

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1   History of CLIL

Immersion programs started with the primary objective to instill bilingualism among students by making the second language (L2) of the learners the medium of instruction for a non-language subject area or content, instead of teaching and learning the second language as a separate entity. The earliest well-known immersion program is the Canadian-French full immersion programs in which monolingual English speaking learners are immersed in a French (L2) instructional environment. In the United States, however, the immersion approach has diversified into many different forms and it is more frequently known as Content-based instruction (CBI). CBI-related programs range from Sheltered English-as-a-second language (ESL classes) for immigrant children, Two-way Immersion (TWI) program for heterogeneous groups of majority-language and minority-language learners, as well as theme-based foreign language courses (Tedick & Wesley, 2015). CBI is different from Canadian-French immersion particularly in its diversity of students' L1 backgrounds and its priorities to help immigrants learn the majority language as well as to preserve Native American languages.

         The term Content and Language Integrated Learning (CLIL) then came into existence in the early 1990s by a group of experts from different backgrounds. CLIL has always associated with the political goal of the European Union (EU) to advance bilingual/ multilingual education as a means to create a multilingual society in Europe due to globalization as well as the ever-increasing demands in the modern day's social and economic forces [Marsh 2002; Cenoz & Ruiz de Zarobe, 2015]. The root of CLIL can always be traced back to the immersion-related programs in Canada and the United

States. From the first introduction in the 90s until today, CLIL is still a popular research area.

1.1.1 Complexity of CLIL in Conceptualization and Implementation

Even though CLIL is no longer a new term in the second language or foreign language education, the exact nature of CLIL, from its definition to the implementation, is still highly contentious among scholars and practitioners due to different theoretical perspectives and conceptualizations about CLIL. Scholars have worked out different ways to conceptualize CLIL programs using different theoretical underpinnings (e.g. Second Language Acquisition theories, sociocultural theory, systemic functional linguistics, etc.). Besides, the declaration about CLIL by the EU commission has often been comprehended in different ways due to a lack of explicit specifications especially for language learning aspects in CLIL. Therefore, when it comes to bringing CLIL into practice, ranging from individual institution, district, region, state to supranational levels, CLIL is diversified and is made even more complex when each school adds its own regulations and provisions (Cenoz, Genesse, & Gorter., 2014).

On a macro level, most scholars would agree on CLIL being essentially an umbrella term for any educational setting or program that involves learning of a target language and a content area [Stroller (2008) in Cenoz (2015); Cenoz et al., 2014]. However, whether CLIL is 'language first' or 'content first', Būdvytytė-Gudienė and Toleikienė (2008) summarizes it well: "the specialists in the didactics of foreign language and subject have not come to an agreement whether CLIL is first of all the development of target language skills, or it is a subject lesson where foreign language becomes a tool for learning another subject" (p.91).

Cited by many, Coyle, Hood and Marsh (2010) [in Cenoz (2015)] refer CLIL as "a dual-focused educational approach in which an additional language is used for

the learning and teaching of both content and language" (p.1); to put simply, CLIL is any dual or joint focus on language and content for instruction (Marsh, 2002). As a result, these definitions allow CLIL to include a wide range of educational practices even if the proportion of language and content is 90 per cent to 10 per cent, which is commonplace in traditional non-CLIL L2/FL education. This resonates with Met's (1999) continuum of content-based language teaching practices with the 'content-driven programs' such as immersion at one end; and the 'language-driven programs' such as conventional language classes at the other (see Figure 1.1). In a similar vein, Paran (2013) also categorizes CLIL into 'weak' and 'strong' CLIL forms: 'weak' CLIL prioritizes language objectives with minimal focus on content while in 'strong' CLIL, content objectives are emphasized alongside complementary language objectives.
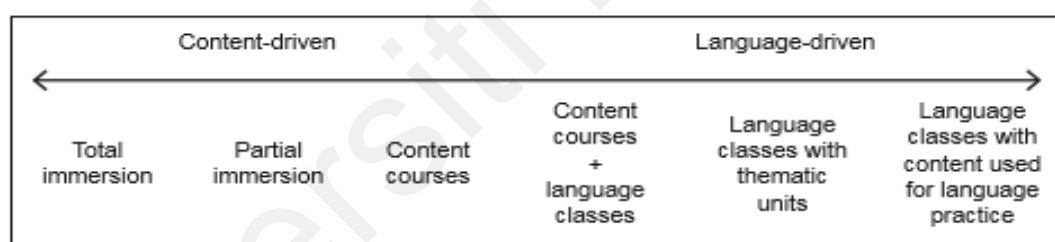
Figure 1.1. Met's [1999, p.41, cited in Lyster & Ballinger (2011)] continuum of content and language integration

These broad and all-inclusive definitions are deemed to be problematic and ambiguous due to its inability to differentiate the hypothetically unique CLIL from conventional L2/FL educational practices which also often involve some extent of integration with content, hence lacking in practical or theoretical utility (Cenoz et al, 2014). However, putting the tug of war between content and language aside, it is helpful to recognize the bottom line of CLIL which is its twofold objective: to ensure that students acquire knowledge of non-language content and to develop their

competence in the target language, and it is not 'only content' nor 'only language' (Eurydice, 2006).

Nevertheless, the most fundamental question of "what is CLIL?" cannot be answered by an over-simplified definition without unpacking at the same time the whole package of CLIL from its goals and characteristics in concept to its actual classroom practices. Any attempt to explain CLIL by quoting one or two definitions cannot do justice to the complexity and enormous depth of the entire CLIL scholarship.

Next, as far as goals and objectives of CLIL are concerned, as a result of multiple theoretical perspectives, it is difficult to articulate the exact goals intended across all CLIL contexts. Other than the aforementioned language proficiency and curricular content area competence concerns, some other CLIL programs also aim for gains in intercultural aspect of learning a foreign language, diversifying teaching methods and classroom practices, cognitive development, self-confidence, learning motivation, plural-literacy, learner autonomy etc. (Dalton-Puffer, 2008). For this study however, the focus of the research mainly pertains to language goals.

Lastly, from a multitude of literature concerning CLIL, it is not difficult to notice the diversity of CLIL in its implementation as well. Dalton-Puffer (2011) identified several characteristics that are considered typical for most CLIL programs in Europe, South America and some parts of Asia: 1. CLIL is about using a foreign language (FL) instead of a second language (L2) so the target language is usually predominantly learned in formal classroom setting rather than students' everyday life in the community. 2. Teachers are normally non-native speakers and they are usually subject teachers instead of foreign language experts. 3. In CLIL programs, typically less than 50% of the curriculum is taught in the target language. 4. CLIL is usually implemented after learners have already acquired literacy skills in their first language

(L1). Apart from that, it is also generally observed that most of the CLIL instructions do not involve explicit language input (De Graaff, Koopman, Anikina, & Westhoff, 2007) and the development of subject knowledge is often prioritized, while a target language is only a vehicle for subject knowledge acquisition [Koch (2005) cited in Būdvytytė-Gudienė and Toleikienė (2008)].

With that said, one important caveat that needs to be added to these prototypical features of CLIL is that not every CLIL conforms to all these characteristics and regulations so the membership of CLIL is irregular, and its structure is fluid and context-dependent. Therefore, within the typology of CLIL itself there are already many variations of implementation with respect to (but not limited to): the balance between language and content instruction, the nature of the target languages involved, instructional goals, characteristics of student participants and pedagogical approaches etc. (Cenoz et al., 2014). In summary, the scope of CLIL is not clear-cut so its core features might differ from context to context. The lack of precision and consistency of its defining characteristics will become more amplified when CLIL is compared with other terms such immersion and content-based instruction (CBI).

1.1.2 Relationship between Immersion, CBI and CLIL

Compared with traditional L2/foreign language teaching, CLIL is distinguished for its heavy focus on content; it is often branded as a "forward-looking and innovative" concept (Dalton-Puffer, Nikula, & Smit, 2010, p.3). At the same time, however, other forms of immersion with different labels also have a long tradition in L2/ foreign language education throughout the world. This raises the question of how exclusive CLIL really is from other types of immersion programs such as Canadian-French immersion, Content-based Instruction (CBI) etc.

In the recent years, though not agreed by all scholars (e.g. Lasagabaster & Sierra, 2010; Dalton-Puffer & Smit, 2013; Pérez-Cañado, 2012), many researchers (Cenoz, 2015; Ruiz de Zarobe, 2008; Marsh 2002) have advocated for the convergence of terms – so whether it is immersion, CBI or CLIL, they can be used interchangeably. This call for integration of terms can be attributed to first, the similarities of essential properties shared among the programs and second, the inconsistent description about the parameters which define each program, therefore distinguishing a program based on its name (CLIL or CBI) without examining its attributes closely might be a superficial attempt.

The most well-known immersion programs are probably the Canadian-French full immersion and one/two-way immersion in the United States. In order to consider a program to be immersive, 50% of the instruction for contents must be taught in the target language (Tedick, Christian, & Fortune, 2011, p.2). Different immersion programs with various "combinations" of content and language elements [see Met's (1999) content-driven and language-driven continuum; Figure 1.1] were later then included as 'content-based second language instruction' (or CBI), an umbrella term that is used in the US education system. From the surface, it is acknowledged by many that CLIL and CBI are indeed different in many ways. However, Cenoz et al. (2014) analyzed the claims about the uniqueness of CLIL and its differences from CBI, and stated that in terms of the essential or defining properties of CBI/CLIL (such as language goals and content-language integration), they are essentially similar despite having different labels. However, with regards to other properties of CBI/CLIL that can vary a lot in different contexts such as starting age, history of origin, L2/FL as the target language, native vs non- native teachers, these are considered as accidental properties because they do not define these programs (Cenoz, 2015; Paran, 2013).

Therefore, it is believed that CBI and CLIL are essentially the same (Cenoz, 2015; Ruiz de Zarobe & Cenoz, 2015; Lyster & Ballinger, 2011; Genesee & Lindholm-Leary, 2013; Ruiz de Zarobe, 2008).

Apart from the shared similarities, another reason that might have contributed to the call for integration of terms is the inconsistencies of characteristics within each label. For instance, the core characteristics of CLIL itself are understood in different ways in terms of the language-content proportion, the nature of the target languages, instructional goals, learner characteristics, classroom compositions, teaching methods etc. In other words, there are no two CLIL or two CBI contexts that are the same.

The most prominent inconsistency within CLIL and CBI is the nature of target language. CLIL has often been associated with foreign language (FL) learning; while CBI is closely tied with L2 learning. Some even claimed that CLIL is strictly catered for foreign language learning but not second languages (Dalton-Puffer & Smit, 2013; Cummins, 1998). However, distinguishing CLIL from CBI merely on the basis of whether it is a foreign or second language is bound to cause confusion because the norm "CLIL is for FL and CBI is for L2" has been belied in many contexts. For example, the "English for Japanese speakers" program in Japan (Bostwick, 1998) is called an immersion program even though English is often considered as a foreign language in the monolingual Japanese society. Besides, in CBI immersion programs, there are many immigrant students who do not have the target language as their L2 (Cenoz, et al., 2014). Lastly, in some European countries like Spain and Sweden where CLIL is implemented, the students receive a lot of exposure to English outside of class (Sylvén, 2013), hence English is not necessarily a foreign language in that context compared to other countries where learners receive target language instruction solely through formal education in school.

It is also important to mention that the adoption of the term CLIL, in the first place, was linked to, according to March (2002), the rejection of the term immersion and "recognition that Europe is not Canada…, (leading) it to a seeking out for alternative terms" (p.57). Therefore, the diversification of CLIL contexts within Europe itself seems to defeat the original purpose of adoption of a new term, at the same time it highlights the shortfall of a terminology to faithfully depict the actual context-sensitive nature of CLIL or CBI.

In conclusion, for this study, concurring with many seminal CLIL researchers' [such as Stroller (2008) and Mechisto, Marsh and Frigols (2008) cited in Cenoz (2015)] view, an integrated outlook of CBI and CLIL will be adopted. First, it is important to maintain a truly international perspective under the umbrella of content and language combined approach, at the same time to put an end to the common but unhelpful exclusivist geographical associations between "CBI and North America" and "CLIL and Europe". This is because CLIL, immersion and CBI share many essential similarities and differences in some aspects (e.g. in geography) are considered non-defining. Second, it is crucial to acknowledge the risk of trusting the 'brand' or labels alone because as shown above, within CBI or CLIL itself contains many inconsistencies (or variations) in implementations so as Ruiz de Zarobe and Cenoz (2015) aptly puts it, "an attempt may be even futile when we talk about the term itself" (p.91) without looking into the educational and contextual variables of a CLIL program.

Having said all that, it is not of the writer's intention to dwell on the nomenclature issue, therefore, for the purpose of brevity, CLIL will be the terminology used throughout the study even though examples from CBI and immersion education might also be drawn. In the next section, the context of CLIL in Malaysia under a

different label called Dual Language Program (DLP) will be discussed along with an overview of English learning culture in Malaysian society.

1.2   Background of Study

1.2.1   Goals of English Learning and Teaching in Malaysia

The English language education in Malaysian public schools has always had communicative competence designated as one of the pillar goals in the national English language curriculum. Nonetheless, in the most recent years, there seems to be a shift of attention to academic English or English for specific purpose when in general students are expected to pursue tertiary education. Academic English entails multiple complex features of English required for learners' long-term academic success and then their career advancement and professional growth (Scarcella, 2003).

Besides, in this modern society driven by knowledge-based economy, more emphasis has been given to Science, Technology, Engineering and Mathematics (STEM) subjects which are believed to be at the heart of modern life and economic prosperity. There is an unprecedented demand for STEM talent and it is believed that world's future jobs are rooted in STEM. English can be said to be the universal language of science and technology because it is predominantly used around the world to communicate information and knowledge (Selamat, Esa, Hj. Saad, & Atim, 2011). In this sense, English is no longer merely a language for international affair and global engagement, it facilitates the acquisition of knowledge (Selamat et al., 2011); along with the rising popularity of STEM education, for Malaysian students to be academically successful in universities and also to build up strong STEM human capital, academic English in particular is therefore indispensable.

1.2.2 Dwindling English Proficiency

Nevertheless, Malaysian students' English proficiency in general has been deteriorating over the years. The poor mastery of English has become a hindrance for Malaysian students to pursue studies in reputable universities overseas (Arukesamy, 2015) and also has taken a toll on the employability of the local graduates. When it comes to tertiary education, students are required read specialized textbooks and academic materials so many students find academic English particularly challenging. Nambiar's (2007) study has revealed that Malaysian tertiary learners are often weak at understanding long sentences or sentences with difficult words and are not able to operate autonomously when they engage with a range of academic reading tasks (Kaur & Sidhu, 2013). Apart from that, local undergraduates are ill-equipped with necessary vocabulary level and size for tertiary studies (Harji, Balakrishnan, Kaur Bhar, & Letchumanan, 2015; Abd. Manan, Ali, & Shamsudin, 2013). In a study conducted by Harji et al. (2015), it showed that almost none of the students had acquired satisfactorily the academic vocabulary listed in University Word List (UWL); in fact, most of the students' vocabulary level was shown to be at 2000-word level according to Nation and Laufer's (1999) Productive Vocabulary Levels Test.

Therefore, it has become clear that prioritizing academic English in school is important to ensure academic success for the students. However, there is a potential negligence of the teaching of academic words in secondary education, according to Abd. Manan, Ali, and Shamsudin (2013). In their study that specifically investigated academic vocabulary learning in Malaysian secondary school education, they found out that out of 1,303 target words to be learned by students from Form One to Five (as listed down in the national syllabus), only seven words belonged to Academic Word List (AWL), a list of words frequently encountered for academic purposes.

1.2.3 Lessons from the Past Program- PPSMI

Recognizing that English language was necessary for the future development of the country, in 2003, the Prime Minister then, Tun Mahathir announced about the implementation of Teaching of Mathematics and Science in English policy (ETeMS, also known as PPSMI in its Malay acronym, which stands for *Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris*). English was then made to replace Malay as the medium of instruction in schools for the teaching of Mathematics, Science and Technology subjects, hoping to address the poor English proficiency issue among Malaysian graduates and also to enable students to tab into the worlds of Science, Mathematics and Technology (Mohd. Saat & Othman, 2010). However, PPSMI policy was retracted in 2012 due to many criticisms incriminating English language for learners' poor Math and Science achievements (Singh, Rahman, & Hoon, 2010). Therefore, the Malaysian government decided to reverse the PPSMI policy back to teaching Math and Science in Malay language.

Nevertheless, looking at the history about PPSMI in retrospect helps articulate and understand CLIL attempt in Malaysian educational climate. Some of the implementation issues that arose during PPSMI are still be relevant to the new CLIL program, namely Dual Language Program (DLP). PPSMI-DLP implementation issues will be juxtaposed in Chapter 5.

1.2.4 Dual Language Program (DLP)

After the reversion of PPSMI, when things seemed to be going back to normal, a nation-wide survey conducted in 2015 showed that 95.7 per cent of the respondents regarded highly the urgency to include more English experience in public schools, hence requesting more subjects to be taught in English to increase proficiency in English among Malaysians (Jala, 2015). On top of that, in the latest Malaysian

Education Blueprint (2015-2025), MBMMBI policy or *Memartabatkan Bahasa Malaysia & Memperkukuh Bahasa Inggeris* (translated as: To uphold Malay language and to strengthen the English language) was introduced to endorse the legitimacy of Malay as a medium of communication in all national schools and secondary schools, at the same time to instill bilingualism (Malay and English) in each child.

Under the umbrella of MBMMBI policy that underscores both Malay language as the national language and English as an international language, in 2016, Dual Language Program (DLP) was introduced by the Malaysian Ministry of Education. Unlike PPSMI, DLP-qualified schools are given the autonomy to use either English or Malay as the medium of instruction for Mathematics and Science; parents are also bestowed options to enroll their children into either DLP or non-DLP classes. The pilot phase started in January 2016, involving 300 primary and secondary schools, and it is believed that the number of participating schools will increase as years go by. More in-depth information about DLP and the selection of DLP students will be discussed in 'Selection of DLP Schools and Students' section in page 47.

As stated in the official circular (KPM, 2015) from the Ministry of Education, the objectives of DLP are to (as translated from Malay): 1. reinforce mastery of English; 2. increase students' access and exploration of knowledge in order to be globally competitive by having strong bilingual background. In simpler terms, given that English is the universal form of communication in Science, DLP aims to increase students' English proficiency as well as their competences in STEM subjects.

DLP fits into the basic description of a CLIL program as stated in the introduction for its conspicuous dual-focus nature: it targets on learning of English (a second language) through Math and Science (content areas). Therefore in this study, 'CLIL' and 'DLP' will be used interchangeable. Lastly, due to the fact that it is one of

DLP's goals is to prepare students for the STEM field driven economy, it is relevant for this study to investigate specifically learning of academic English in DLP classroom.

1.3   Statement of Problem

Amongst the many research interests pertaining to CLIL, the focus has always been predominantly on L2/FL language learning and its outcomes. While most of the empirical studies are leaning towards the positive contribution of CLIL towards language learning in different aspects [e.g. speaking and listening in Arribas (2016); general language competence in Mattheoudakis, Alexiou, & Evosmos, (2012)], however, there are also studies (Dallinger, Jonkmann, Hollm, & Fiege, 2016a; Pladevall-Ballester & Vallbona, 2016; Sylvén, 2013) that point out that the results of CLIL are not necessarily positive in all contexts and that CLIL does not always warrant better learning results compared to traditional L2/FL learning classrooms. For instance, amongst studies about CLIL students' vocabulary learning, the positive outcomes (e.g. Merikivi & Pietilä, 2014; Moghadam & Fatemipour, 2014; Alonso, 2015) are strongly counteracted by studies that presented either negative or non-significant evidence of CLIL effect towards students' lexical development (e.g. Agustín-Llach , 2015); Arribas , 2016; Tragant, Marsol, Serrano, & Llanes, 2016).

This apparent divide in the literature could be attributed to the context-sensitive nature of CLIL. The implementation of a CLIL program is highly nuanced due to various factors ranging from institutional to individual learner levels (such as institutional policies framework, teachers' quality, students' cognitive abilities etc.), resulting in diversification of CLIL settings. For instance, commenting on general European CLIL landscape, Sylvén (2013) pointed out the discrepancies among European countries such that CLIL in one country is not necessarily the same as in

another. Therefore it is important to take careful consideration of nation-specific contextual factors which influence learning in different countries (Meriki & Pietilä, 2014) because a CLIL program might work well in one individual situation but not the other due to its highly context-sensitive nature.

To intensify the debate on the learning outcomes of CLIL, in the more recent studies (Llinares, 2015; Dallinger et al., 2016a), some scholars even claimed that the positive effect of CLIL on language learning in the past were overestimated and its shortcomings were under-addressed. Baneges (2011) termed this phenomenon as the "bandwagon effect". This apparent skepticism towards the novelty of CLIL can be attributed to the methodological flaws in some of the past studies.

First of all, as claimed by Dalton-Puffer et al. (2010), CLIL is often regarded as the alternative to conventional language learning because of its 'innovative, modern, effective, efficient and forward-looking outlook' (p. 3). Therefore, it has a high tendency to attract students who have better and higher prior academic achievement, general cognitive abilities, motivation and demographic background, hence expectedly producing better outcomes which might be misconstrued as CLIL effect (Lorenzo, Casal, & Moore 2010, Bruton, 2011, 2013; Dallinger et al., 2016a). For many studies that compared CLIL and non-CLIL group, this selection effect was not controlled, therefore, it is difficult to establish a causal relationship between CLIL and learners' language outcomes (Coyle, 2007; Bruton, 2011). Apart from the selection effect, when interpreting CLIL-effects in CLIL/non-CLIL comparison, other factors need to be controlled for as well. For instance, Sylvén's (2004) study showed that it was the extramural exposure to L2/FL outside of school that contributed to the higher performance of CLIL students rather than CLIL instruction in the classroom per se. Besides, CLIL instruction usually entails more contact hours with the target language

during the school day, and it could be the extra hours of target language lesson that results to better performance among CLIL students (Bruton, 2013; Dallinger et al., 2016a; Tedick & Cammarata, 2012). This raises doubt on the necessity of having CLIL at all because given the same number of hours of conventional language instruction, it would probably be as effective (Bruton 2011). Therefore, in any comparative study, the instructional hours of two groups should be controlled to ensure that the effect of increased exposure time to target language is not confounded with CLIL effect. Failure to control some of these factors raises doubt to the validity of the research methodology and hence the generalization resulted. In sum, these methodological issues and concerns (e.g. the lack of control over CLIL selection effect, total formal instruction hours in English and extra-curricular exposure to English) have prompted many researchers to call for the reevaluation of the past seemingly over-optimistic CLIL results in order to discern the true academic gains as the result of CLIL practice" (Pérez Cañado, 2012).

Lastly, there is a dearth of information on CLIL and academic language. One of the important components of academic language is academic vocabulary. Taken altogether, even though there are relatively many research done in the area of general vocabulary, due to the aforementioned inconsistent language outcomes, the context-sensitive nature of CLIL and the skepticism towards past positive results, the generalization of CLIL as a foreign/ L2 language teaching approach is difficult. Therefore this study aspires to fill these knowledge gaps by looking into Dual Language Program (DLP), a CLIL program in Malaysia. DLP is perceivably different from other CLIL programs in commonly studied regions like North America and Europe due to its unique context, especially in terms of students' L1 diversity, teacher training in language and content integration, and DLP policy in the national education

system. Therefore, it would be interesting to look at the CLIL landscape in Malaysia in order to gauge the success of CLIL approach to improve Malaysian students' academic English learning at the same time to overcome the methodological flaws in previous studies that might be skewing the actual CLIL effect. It is hoped that a study that is sensitive to the overtone of the context like this could help add breadth to the CLIL scholarship and as Ruiz de Zarobe and Cenoz (2015) put forth in their recent review article, "to stimulate comparisons to reach some conceptual and pedagogical generalizations" (p. 94) about CLIL itself.

1.4   Research Objective

This research is comparative in nature that compares the academic vocabulary performance between CLIL and non-CLIL groups. The main objective of this study is to find out if students in two different learning environments, i.e. regular mainstream classes and CLIL (Content and Language Integrated Learning) instruction, perform differently in English academic vocabulary test, given three important predictors as shown in the literature – selection effect, total instruction hours in English in schools and extramural exposure to English (see 'Call for Critical Assessment of CLIL', p.41) are controlled.

1.5   Research Questions

To achieve the objective above, this study is guided by three research questions:

- Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Two non-DLP groups given that the selection effect is controlled?

- Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Four non-DLP groups when both groups' total formal instruction hours in English are controlled?

- Is there any difference between DLP and non-DLP students' Extramural English Exposure? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English a confounding factor in this study?

1.6   Theoretical Framework

English is recognized worldwide for its importance and ubiquity in many disciplines, from informal day-to-day communication to formal and cognitively demanding subject matters such as business, science, law, etc. Many second and foreign language programs aspire their students to not only be competent in social language, but also proficient in academic settings. A content-language integrated approach is believed by many to be a fertile ground for developing academic language (Būdvtytė-Gudienė & Toleikienė, 2015; Lorenzo & Rodrígues, 2014; Anderson, 2011). In order to investigate if CLIL leads to the better learning of academic language, the theoretical basis for this study regarding academic language has to be first established.

The construct of academic language hinges on the distinction of two proficiencies that Cummins (1981) made: Cognitive Academic Language Proficiency (CALP) and Basic Interpersonal Communication Skills (BICS). It was motivated by Cummins' intention to highlight challenges immigrant second language learners face as they attempt to catch up to grade-appropriate academic aspects of language (Cummins, 2008). The BICS/CALP distinction is deemed necessary because misconception about the nature of language proficiency has detrimental implications on language learners' intellectual experiences and participation in academic tasks (Cummins, 1999). Cummins (1980 &1981) showed that in bilingual contexts, many English learners experienced academic difficulties; when administered the cognitive ability test, their development rate appeared to be slower than normal so they were

'diagnosed' as learning disabled and inappropriately placed in special education programs while it was actually the lack of academic language proficiency that caused the students to perform poorly academically. This is the result of the wrong assumption that children who could converse effortlessly in the language would not have problem in learning with that language. Therefore, BICS/CALP distinction is vital to unleash the academic potential of L2/FL students who have the target language as the medium of instruction.

1.6.1 Cognitive Academic Language Proficiency (CALP)

Cummins (1999; 2008) formalized the differences between BICS/CALP as conceptually unique components of the construct of "language proficiency" in terms of acquisition and developmental patterns. Proper understanding of BICS/ CALP proficiency enables practitioners and policy-makers to set precise goals for target language learners to attain not only peer appropriate conversational fluency but also grade-fitting academic language. BICS or conversational fluency is commonly known as language skills needed in daily social situations (e.g. interactions that take place on the playground, during recess time, etc.). On the other hand, CALP or academic language proficiency refers to students' ability to understand and express, in both oral and written modes, concepts and ideas that are relevant to success in school (Cummins, 1999; Krashen and Brown, 2007; Scarcella, 2003). To properly understand what CALP entails, it is best done through comparing and contrasting CALP with BICS.

First of all, Cummins (1982) cited in Roessingh (2005) uses an iceberg metaphor to draw distinction between BICS/CALP proficiency. The surface part of an iceberg resembles BICS which represents only about 10% of the overall proficiency of an academically competent learner while the underneath the surface is the vast underlying proficiency, CALP. Therefore, it is generally assumed that academic

language proficiency requires a longer time to develop than conversational aspects of proficiency (Cummins, 1981). Unlike BICS which normally reaches a plateau in the first six or so years of learning, CALP continues to develop throughout schooling and even one's lifetime (Cummins, 1999).

Other than the acquisition time, Cummins also believes that both BICS and CALPs are distinct with respect to the contextual support and cognitive demand in linguistic activities or tasks. Hence his framework for the development of language proficiency [see Figure 1.2] elaborates BICS and CALP based on two intersecting continua: context-embedded to context-reduced (horizontal axis) and cognitively undemanding to cognitively demanding (vertical axis).

Cognitively undemanding

• survival "chunks" • simple grammar forms • high- frequency vocabulary: family, clothes, food, money, face-to-face interactions. • "here and now" language: 1000-2500 words. 1	• initial reading skills • writing for personal needs: notes, lists, recipes, group-constructed text • common vocabulary: sports, hobbies, celebrations • begin to integrate grammar and vocabulary: mine-themes • "my lived experienced": 2500-5000 words 2

Context embedded ——————————————————→ **Context reduced**

3 • transitioning to curriculum-related content. • shift from learning to read to reading to learn • thematic units • ESL learn has 3000 high-frequency words, some academic vocabulary (AWL) and some common vocabulary. • 6000 words	4 • "the educated imagination": ideas I can access only through language itself • abstract thought • extensive use of reading and writing on academic genres (essays, debates) • 12,000 words +

Cognitively demanding

Figure 1.2. Cummins' Framework of BICS/CALP modified by Roessingh (2006), p.93

Located in the first quadrant, BICS or the language of "here and now" and "my lived experience", often appears in everyday social interactions which are usually context embedded; therefore, meaning is easily conveyed using paralinguistic cues such as body language, intonation, facial expression etc. Besides, even when contextual clues are reduced as shown in quadrant 2, it is not very demanding

cognitively; in other words, higher order thinking is less involved and precise understanding is seldom required. Apart from that, BICS often involves high frequency vocabulary so to become proficient in BICS, language learners experience relatively less difficulty than learning CALP.

On the other hand, CALP in quadrant 3 and 4 are more decontextualized (clues are language clues unlike BICS). As shown in quadrant 3, learners begin to transition into curriculum-related content and from 'learning to read' to 'reading to learn'; and in quadrant 4, it is characterized as the acquisition of metaphoric competence. Both quadrant 3 and 4 are cognitively demanding because new ideas, abstract concepts and complex language structures are presented to the students at the same time. They also involve usage of technical vocabulary, academic words and less frequently used vocabulary. Due to its cognitively demanding nature, knowledge of CALP is believed to be vital in gaining academic success.

1.6.2   Criticism and Rebuttal

Cummins' BICS/CALP distinction was heavily criticized by Scarcella (2003) in her conceptual framework for academic English. First, she refers the BICS/CALP distinction as the dichotomous conceptualization of language that fails to depict the complex reality of academic language that involves many variables. This has led many to believe that the development of BICS or CALP is monolithic and can be achieved in separation when in fact, the development can be of both CALP and BICS. The example Scarcella gave is the teaching of learners to hear and identify phonemes that not only facilitates the development of basic oral proficiency, at the same time it contributes to the development of CALP because it helps learners to access academic vocabulary when doing advanced reading.

Second, Scarcella finds the BICS/CALP distinction of limited practical value. As Cummins' notion of BICS and CALP comprises a degree of cognitive challenge, a simple linguistic dichotomy that merely distinguishes between oral and literate forms of language is insufficient to operationalize tasks and therefore does not generate tasks that teachers can use to help develop their students' academic English.

In response to Scarcella's criticism that claims BICS/CALP notion as being oversimplified, in Cummins's (2008) rebuttal, he contends that such criticism derives from taking the constructs out of their original context. The BICS/CALP distinction was initially formulated to address certain theoretical issues that Oller (1979) proposed (e.g. whether "language proficiency" could legitimately be viewed as a unitary construct) and was not formulated as a tool to generate academic tasks in the first place. Furthermore, the distinction was not proposed as an overall theory of language, but a conceptual distinction; therefore, there was never any suggestion that these were the only important components of that construct. (Cummins, 1999, 2008).

Cummins (1999) made it clear that to say that BICS and CALP are conceptually distinct is not the same as saying that they are separate or acquired in different (binary) ways. This is because beginning from home, children in general build up their conceptual foundation (understanding of the word) through conversing with family members. In the same vein, cognitive skills are necessitated in all kinds of social interaction to a certain degree. Therefore, this intersection of the cognitive and social aspects of language proficiency does not mean that they are interchangeable, neither are they mutually exclusive.

1.6.3 Academic English

English is not only one of the most spoken languages, English is also the internationally recognized language in numerous fields such as economics, commerce,

politics, science, technology, academics etc. Academic institutions worldwide would have no problem acknowledging that English is the most prominent language in achieving academic purposes. Even though Cummins (1999) did lay out three components of CALP construct, as Cummins himself has emphasized, it is just a conceptual distinction instead of a whole theory. The three components are: 1. Cognitive (instruction should involve higher-order thinking abilities rather than the low-level memorization and learning by drilling) 2. Academic (content should be integrated with language instruction as in content-based ESL programs) 3. Language (the development of critical language awareness). Due to the fact that the BICS/CALP distinction was not intended for creating academic tasks, hence, Cummins (2008) recognized the usefulness and suitability of Scarcella's framework that details various dimensions and components of academic English for academic task generation purposes.

Therefore, in order to articulate what it means to be proficient in academic English for the purpose of this study, Cummins's BICS/CALP distinction serves as the theoretical backdrop insofar as language proficiency is concerned and as the BICS/CALP notion is purely conceptual (Cummins, 1999), it will be supplemented with Scarcella's (2003) framework for academic English.

Academic English is broadly defined as a genre of English that is used in professional books and is characterized by the specific linguistic features associated with academic disciplines (Scarcella, 2003); in other words, it can be called as the "language of schooling" (Schleppegrell, 2001). This is the reason the mastery of academic language is highly associated with academic success for English learners and has always been referred to as the key to content area learning (Bailey, Butler, & Sato, 2007; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006). Some of the academic

English tasks include taking notes to a lecture, reading content class textbooks, presenting a science experiment report, taking standardized tests and many more. Scarcella discusses many various dimensions and components of academic English. First of all, just like the everyday communication, in academic situations, it involves four language skills—reading, writing, speaking, and listening and also other language components. However, academic language is different from the social language for its distinctive vocabulary, syntax, and discourse features that are commonly used in advanced reading and writing required by grade-level curriculum (Bailey & Huang, 2011). Apart from the linguistic aspect, academic English also encompasses cognitive aspects (such as students' background knowledge and higher order thinking, strategic and metalinguistic awareness of language etc.) that are less prominent in social English.

To sum up this section, distinguishing CALP from BICS is indispensable and relevant to the CLIL program in Malaysia because one of the goals of English language education in Malaysia is to prepare learners for the globalized "knowledge society"; hence the development of CALP is essential. CALP or academic English has many facets. As for this study, it is limited in scope to academic vocabulary, one of the linguistic dimensions of academic English. Academic vocabulary will be further discussed in the next chapter (see 'CALP and Academic Vocabulary', p.33).

1.7   Significance of Study

The significance of this study will be discussed from two aspects: international CLIL comparative purposes as well as national-level needs. Firstly, a study that strives to be context-sensitive and methodological careful like this is hoped to add on to an existing empirical foundation in the field by challenging or affirming existing claims about CLIL (see 'Problem Statement', p.13) which might be useful in the making generalizations about CLIL as an approach in the future.

On the other hand, in distinguishing social language (BICS) and academic language (CALP), Cummins (2008) indicates that it requires more time to be competent in academic language, thus even though students might speak fluent English on the surface, CALP is needed for academic success. Abd Manan et. al.'s (2013) study that analyzed the percentage of English academic vocabulary in Malaysian secondary textbooks proved that there is a severe lack of academic vocabulary teaching in Malaysian schools. Students might be ill-prepared for a sudden increase in difficulty of vocabulary used when they transition from secondary to tertiary studies. Therefore, studies about academic language learning, specifically academic vocabulary are urgently needed since there is an increasing number of Malaysian students entering tertiary education today. In other countries, CLIL approach seems to offer a fertile ground for academic language learning (e.g. Lorenzo & Rodrígues, 2014). Therefore the implementation of CLIL in Malaysia appears to provide a solution to address the issue of limited academic vocabulary knowledge prior to tertiary education.

However, there is a dearth of information and empirical research about CLIL in Malaysia, even more so for CALP-related CLIL studies. This study does not just present CLIL-non-CLIL comparison in superficial manner, it seeks to expound the results by looking into some highly contextualized implementation issues. Therefore, this study could be used as one of the reference points for different stakeholders to help improve practice and policy despite certain generalizability limitations of the research design.

1.8   Operational Definitions

- CLIL/ non-CLIL: CLIL is an umbrella term, which refers to a dual-focused educational context in which a non-first language is used as a medium in the teaching and learning of non-language content (Marsh, 2002). However, as

mentioned in the introduction, CLIL is often understood in different ways as there is neither one CLIL approach nor one theory of CLIL. Therefore, it is necessary to admit that CLIL may, and does, take any form or shape in different geographical contexts.

CLIL in Malaysia refers to an educational context that has a dual-focus in both content and language even though admittedly Malaysian CLIL environment is content-oriented while language learning is believed to happen naturally with no explicit language objective. On the other hand, non-CLIL or traditional or mainstream classroom refers to an educational setting that sees content and language as separate entity.

- CALP: The terms *academic language proficiency* and *academic language* are used interchangeably with CALP (Cummins, 2008). CALP is distinguished from BICS by Cummins (1979) to reflect the language that children acquire in school to understand and express in both oral and written modes concepts and ideas that are relevant to progress successfully through the grades school (Cummins, 2008). There are many facets of CALP (e.g. linguistics, cognitive etc.), but this study merely focuses on the lexical aspect.

- Selection effect: Even though CLIL claims to be open to all learners, voluntary participation does not always guarantee a spot in CLIL classroom because when there is a surplus of applications or other logistical predicaments, schools select CLIL candidates based on their perception of likeliness to succeed in CLIL (Dallinger et al., 2016a). Therefore, regardless of the implicit or explicit nature of the selection process, the selection mechanism itself has unavoidably favored students with certain characteristics (Bruton 2011, 2013) and this phenomenon is referred as the selection effect. As shown in another study by

Dallinger, Jonkmann, and Hollm's (2016b), it indicates a strong selectivity of CLIL programs such that at the onset of lessons, CLIL students already possess more favorable learning prerequisites and advantages compared to their traditional mainstream counterparts in terms of their cognitive abilities, socio-economic background, motivation and prior academic achievement.

- Total instruction hours: Studies such as Alonso (2015) and Pladevall-Ballester and Vallbona (2016) CLIL addressed the comparability issue concerning instruction hours by keeping the amount of formal instruction to English in school between CLIL and non-CLIL groups the same up to the time of testing. As for this current study, the total instruction hours of CLIL and non-CLIL groups were tallied based on these assumptions: (1) similar English learning time in primary school; (2) average of 6-7 lessons of English, Science and Mathematics per week; (3) 30 minutes each lesson.

- Extramural Exposure to English: 'Extramural English' is a term introduced by Sundqvist (2009) in her doctoral thesis which in essence refers to any type of situation that learners come in contact with outside of the classroom that gives rise to English learning opportunities. There are many ways learners are believed to "pick up" new words, such as reading English books, checking web-based material in English, and many more unmentioned (Olsson & Sylvén, 2015; Pietilä and Merikivi, 2014). This study is interested to investigate the types and amount of exposure to English students have outside of school that might contribute to academic vocabulary learning, then confounded as CLIL effect.

- Academic vocabulary performance: Baumann and Graves (2010) broadly define academic vocabulary as vocabulary that occurs more often in academic

contexts than in other contexts. A commonly used list of general academic vocabulary is the Academic Word List (AWL) compiled by Coxhead (2000), which consists of 570 word families extracted from a corpus of 3.5 million running words in academic books and journals of different domains. For this study, learners' academic vocabulary performance was measured using the AWL section of the Vocabulary Levels Test (VLT) (Schmitt, Schmitt, & Clapham, 2001). The higher the test score of a learner, the better his or her academic vocabulary performance is assumed.

1.9   Scope

This study confines itself to comparing secondary CLIL and non-CLIL students in Selangor secondary schools, therefore is not generalizable to all CLIL schools in Malaysia. Besides, in some schools, there are CLIL participants whose L1 is English and are already proficient in English, hence this study is only generalizable to the non-English speaking CLIL students. Apart from that, as it was discussed that there are many aspects of CALP, this study only focuses on students' performance in academic vocabulary. To narrow the scope down even more, this study is only concerned with the receptive academic vocabulary knowledge (ability to understand a word when it is heard or seen) instead of productive vocabulary (ability to produce a word when one writes or speaks). Receptive academic vocabulary is prioritized in this study because it is generally believed that words are known receptively first and L2 learners' receptive vocabulary size is typically regarded to be wider than their productive vocabulary size (Laufer, 1998). Learners' receptive academic vocabulary is measured using Vocabulary Levels Test (see 'Instrument 1: Academic Vocabulary Test', p.67). It is also worth-mentioning that this academic vocabulary test only tests on the breadth

(superficial understanding) of learners' vocabulary knowledge rather than the depth which includes collocation and appropriateness (Merikivi & Pietilä, 2014).

1.10 Conclusion

The introduction chapter starts off with a substantial amount of discussion about the history of CLIL, its complex nature and the general properties of a CLIL program. The chapter then proceeds to the most pivotal part, the problem of the study which addresses several gaps in the literature, giving rise to three research questions. Cummins' CALP/BICS distinction of language proficiency is adopted as the theoretical framework in answering these research questions due to the fact that academic language proficiency and content teaching in CLIL are inseparable. Lastly, the significance of the study, definitions of terms and limitations of this study are also discussed. The following chapter will review some of the seminal studies pertaining to academic language learning, especially in the area of vocabulary, as well as the growing controversy amid the CLIL scholarship.

CHAPTER 2

LITERATURE REVIEW

2.1   Introduction

In Chapter 1, it has been explained that BICS/CALP distinction serves as a theoretical foundation that links CLIL and language learning. CALP is in general more cognitive demanding and it involves more complicated linguistic features. Therefore, it is important for language learners to have high academic language proficiency because the mastery of CALP is believed to be the key to academic success and to achieve the goal, CLIL is deemed to be the promising 'solution' to boost learners' CALP acquisition. However, this does not go without criticism and the call for methodological reevaluation of the past studies. This chapter will begin by discussing about: 1) the rationales of CLIL with regards to CALP acquisition, 2) the relation between academic vocabulary and CALP, 3) past CLIL research on general or academic vocabulary learning, and lastly 4) methodological flaws in previous studies that might potentially affect how CLIL-effect is interpreted.

2.2   CALP Learning through CLIL Instruction

The existing extensive body of research has shown that language development and cognitive development are interrelated and inseparable (e.g. Snow, Met, & Genesee, 1989; Deák, 2014). It is even more so when Cummins distinguishes CALP proficiency from BICS proficiency for its higher cognitive demand to acquire it, thus calling for an integrated approach that brings language and cognitive domains together in instruction. Traditional L2/FL teaching and learning often dissociates language learning from cognitive or academic development; while CLIL is believed by many to be an incubator for CALP as the target language learning is embedded in rich content, as opposed to having language learning and subject matter learning as separate

activities. Therefore, this section is dedicated to investigate the so-called "CLIL-effect" and rationales as of why CLIL instruction is believed to be more effective than other instructions (that do not integrate language and content) in helping students acquire CALP from the Second Language Acquisition (SLA) standpoint.

First of all, Krashen and Terrell's Natural Approach (1983) suggest that language learning can take place effortlessly without formal instruction; in the same vein, Krashen's (1982) Acquisition-Learning hypothesis in the theory of Second Language Acquisition purports that the acquisition of language takes place in 'natural communication' where learners are concerned not so much of the correctness and form of their utterances, taking the pressure off the learners, because it is more of messages and meanings they're conveying that matter most. Therefore a content-language integrated CLIL classroom is believed to be able to create that kind of stress-free environment for naturalistic language learning to take place. The emphasis on the 'meaningfulness' of the interactions in the target language is often lacking in traditional language lessons where inputs are not always as authentic (hence not meaningful to the learners) because language is perceived as a system of linguistic elements that can be anatomized and memorized (Muñoz, 2007).

Along with the idea of meaningfulness, Krashen's input hypothesis calls for the provision of 'comprehensible input' which is input that is one step beyond a learner's current linguistic abilities (i+1) (Krashen, 1982). It claims that fluency in speaking or writing in a second language will naturally happen after learners have built up sufficient competence through comprehending input; CLIL classrooms can provide abundant comprehensible input by incorporating content into language. This is because content provides cognitive basis for language learning such that it provides real meaning, an inherent feature of naturalistic language learning (Snow, Met, &

Genesee, 1989). If students are introduced to new and complex linguistic systems through academic subject content, meaning and language will be connected and thus learning will be enhanced (Genesee & Lindholm-Leary, 2013). In other words, CLIL provides meaningful contexts and hence, comprehensible input for language learning to take place by using authentic texts made possible by subject areas (Būdvytytė-Gudienė & Toleikienė, 2008), activating students' prior knowledge, participating in hands-on activities of different topics (Cummins, 1998) and many more activities in CLIL environment.

Lastly, Krashen believes that affective variables such as motivation, self-confidence and anxiety can either facilitate or impede learners' second language acquisition. For instance, learners who experience high level of anxiety in the acquisition process might not be able to perform well due to the raise of their 'affective filter' or 'mental block'. CLIL has the potential to create a comfortable and non-tense environment because when language learning is blended into subject area instruction, learners will not feel being forced into using the target language; instead, there is a natural transition into producing output as the result of receiving substantial amount of comprehensible input (e.g. students preparing Science lab report after conducting an experiment). Besides, learning language through a content serves as a motivational basis of the whole language learning experience. Without the authentic purpose of learning a language, learners might not feel the need to pay attention to the language. For school children, one of the most prominent things they know and need to know more about is the subject matter of school. Hence, content provides a primary motivational incentive for language learning as it is purposeful (e.g. it provides access to certain topics that are interesting to them or helps them score better academically) and therefore worth-learning.

Apart from Krashen's second language acquisition theory, the benefit of CLIL in CALP acquisition can also be explored using Swain's (1985) Output Hypothesis which argues that the process of producing a linguistic output can benefit learners' language learning when learners notice their gap in knowledge of the L2. However, in traditional language classrooms, the linguistic production of learners is often either limited. Pertaining to this weakness, CLIL is seen as the solution that provides more opportunities for learners to produce linguistic output and facilitates language processing by having meaningful content (Muñoz, 2007). For instance, when it comes to speaking and writing, by incorporating the subject matters that suit the needs and interests of the learners, learners have more authentic opportunities to speak and write which then help consolidate their language learning. Unlike repeating a dialogues in language textbooks or writing a report to an imaginary party where the academic language learning is generally abstract and trivial, in CLIL classrooms, language and content learning are believed to happen simultaneously, so CALP learning opportunities are ubiquitous and unpretentious (Lorenzo & Rodríguez, 2014).

All in all, some of the SLA theories from scholars like Krashen and Swain have impacted how acquisition of L2/FL and even CALP is perceived in CLIL environment. Taken all things together, it should rightly be assumed that learners who receive CLIL instruction should acquire CALP better than their non-CLIL counterpart and same thing holds true when it comes to learning academic vocabulary. Nevertheless, there are many other factors that affect general vocabulary acquisition, not necessarily CLIL per se. Therefore it is important to review some of the literature on vocabulary learning hoping to get a more 'distilled' understanding of CLIL's effect on academic vocabulary. Besides, it is necessary to acknowledge an observation of obvious polarization of CLIL results. While many empirical studies have shed positive light on

CLIL, there are also criticisms and negative evidence against CLIL. Therefore, as how Merikivi and Pietilä (2014) have well put: "Results of this kind (negative result) are in the minority but they certainly warrant a careful consideration of individual CLIL situations…and generalizations should be made with caution" (p. 489).

2.3 CALP and Academic Vocabulary

Having established how learning language through content can be beneficial for CALP acquisition, this section is dedicated to specifically look at academic vocabulary, one of the many facets of academic language. Academic vocabulary is often seen as means for communicating and thinking about disciplinary content (Nagy & Townsend, 2012) and studies have found that vocabulary knowledge is key to effective academic reading comprehension (Qian, 2002; Carlo, August, & McLaughlin, 2004).

The BICS/CALP distinction drawn by Cummins (1979, 1980, 1981) renders that these two types of language proficiencies are different such that BICS is used for everyday social conversation while CALP is the language to ensure success in school for it involves students' abilities to read and express abstract and complex ideas that without rich contextual support. Cummins (1999, 2008) himself explicates that the BICS/CALP notion is essentially a conceptual distinction instead of a complete theory of language so it is never intended to be the only important component in language learning. Therefore, under the umbrella of the BICS/CALP distinction, it is made possible many different ways to conceptualize CALP by different scholars, but there are similarities, especially pertaining to the importance of vocabulary in academic language.

Many would agree that a key component of academic language is academic vocabulary (Taboada & Rutherford, 2011; Scarcella, 2003; Clegg, 2007). Just as to communicate in everyday situations where learners must have the knowledge of

certain vocabulary that is used in various daily activities, similarly, to thrive in academic situations, it is necessary for them to know to a certain extent academic words that are often used in academic settings. Academic vocabulary consists of: (1) technical or content-specific words for specific academic disciplines, (2) nontechnical or cross-discipline words that are specific to school learning across all academic fields. As for this study, the focus is only on the cross-discipline or general academic words because unlike content-specific vocabulary which is often explicitly taught in the classroom, general academic vocabulary is easily overlooked by either language or subject teachers (Clegg, 2007).

Nagy and Anderson (1984) claim that there may be as much as 180,000 vocabulary words that are used in school English. Even though it might seem counter-intuitive, it is important to make clear that not every vocabulary used for CALP/ academic purpose should be considered as academic vocabulary because academic texts also usually involve the 2000 most frequently used vocabulary [e.g. General Service List (West, 1953) that is not exclusive to CALP, but also appear frequently in conversations/ BICS. Therefore, there are words that are overlapped in both CALP and BICS settings. When it comes to word frequency analysis, [see Schmitt, Schmitt, & Clapham, 2001], academic words are generally of lower frequency words (Cummins, 2008).

Hence, it is of the writer's intent to make explicit that for a word to be considered as an academic word, it is not just because it is used in school or it is a low-frequency word, but because it is specific for academic purposes and it occurs frequently in a range of academic subjects. With the CALP/BICS distinction as the theoretical basis of this study, there draws the distinction between academic

vocabulary and general vocabulary, which then is determinant to the selection of instrument of this study (see 'Academic Word List (AWL)', p.67).

2.4 Past Studies about CLIL

2.4.1 Positive Outlook of CLIL and Vocabulary

Considering that CLIL has been what Marsh and Frigols (2007) labeled as "a catalyst for change in language education" (p.33), it is unsurprising that most of the research about CLIL revolves around the target language learning. Most CLIL research were presented with positive outcomes when language performances were compared between CLIL and conventional ESL/EFL classes such that CLIL students were shown to outperform their non-CLIL counterparts in many ways (e.g. Dalton-Puffer, 2008; Coyle, 2008; Marsh & Frigols, 2007). Out of the many of aspects in a language system, one of the great gains acquired by CLIL students is the vocabulary. Even though the focus of this study is on academic vocabulary, studies on CLIL and general vocabulary are of the same degree of relevance because academic vocabulary falls under the Vocabulary umbrella and it is also important to be cognizant of methodological drawbacks in the previous studies to improve the current research design for a more robust study.

Many researchers concur that CLIL students have larger general receptive and productive lexicon and know more lower-frequency words as opposed to students who go through mainstream language classrooms [See review by Dalton-Puffer (2011)]. Four studies of different geographical contexts below maintain the benefit of CLIL towards vocabulary learning.

To begin with, Merikivi and Pietilä (2014) compared vocabulary acquisition between CLIL and regular mainstream language classes in Finland. Vocabulary Levels Test (VLT) by Nation (1983, 1990) and Productive Vocabulary Levels Test (Laufer &

Nation, 1999) were used to gauge the sixth- and ninth-grade students' vocabulary size. In both receptive and productive vocabulary tests, the scores of CLIL students were significantly higher than the regular mainstream group, showing CLIL students' superiority for having broader receptive and productive vocabularies than non-CLIL group.

Besides, Moghadam and Fatemipour (2014) did a comparison of vocabulary development between two groups of secondary students from two schools in Iran. Even though the students were of different schools, the subjects were homogenous in terms of their age (13-14 years old), nationality, gender, and also their prior language knowledge. The researchers claimed that the CLIL students demonstrated better performance in the vocabulary test and thus, a higher ability to retain vocabulary than ordinary students because of CLIL method and its textbooks.

Apart from that, Xanthou (2011) examined content and L2 vocabulary learning amongst primary school students in Cyprus where Science was taught in L2 (CLIL) and L1 (non-CLIL). The experimental and control groups were comparable in terms of their age, prior experience with L2 English, first language, as well as the quality of their teachers; gender and academic/language performance were matched to ensure the comparability of the groups. The experimental groups received three 80-minute Science lessons in L2 English while the control groups received the same lessons in L1 Greek. It was found out that CLIL group outperformed their counterpart on L2 vocabulary knowledge by showing greater improvement in pre-post vocabulary breadth assessments.

Lastly in Spain, Alonso (2015) involved sixth grade primary CLIL, sixth grade primary non-CLIL and tenth grade secondary non-CLIL in his cross-sectional study on students' receptive vocabulary size in CLIL and non-CLIL settings. The results of

these three groups were compared in three combinations: (1) 6th grade primary CLIL and 6th grade primary non-CLIL; (2) 6th grade primary CLIL and 10th grade secondary non-CLIL; (3) 6th grade primary non-CLIL and 10th grade secondary non-CLIL. First, the 6th grade CLIL and 6th grade non-CLIL comparison showed significant differences. However, the comparison between 6th grade primary CLIL and 10th grade secondary non-CLIL showed no significant differences in receptive vocabulary size. Nonetheless, the researcher believe that CLIL is beneficial for vocabulary acquisition since 6th graders could score equally well as the 10th graders. Alonso's method of pairing students of different grades (Grade 6 and 10) to ensure comparability of instruction hours between CLIL and non-CLIL groups will be adopted in this study as well (see Selection of Subjects, p.57).

Apart from the general vocabulary, the specific advantage of CLIL learners seemed to lie in academic vocabulary and words from the 5,000 and above frequency range (low frequency vocabulary) (Dalton-Puffer, 2011). In Lorenzo and Rodríguez's (2014) study about CALP development in secondary CLIL History class in Andalusia (Spain), 244 pieces of narratives about a historical event written by 9th to 12th graders from 4 schools secondary schools were analyzed. A language software, Synlex was used to analyze syntactic complexity and lexical complexity of the texts produced by CLIL students. Specifically in the area of vocabulary, the complexity and variation of vocabulary used in the narratives were shown to be consolidated in higher CLIL-grade, hence confirming the potential of CALP development in CLIL setting. The purpose of this study was mainly to observe how CALP structures (syntax and lexicon) changed at different grades of CLIL, but without having a comparison group it could not be used as an evidence to support the claim that CLIL is better than conventional language

classroom. Nevertheless, it is important as a testimonial ratifying CALP learning in a CLIL setting.

The positive effect of CLIL towards academic vocabulary acquisition can be found in the three studies below. Admittedly there are not as many studies that investigate academic vocabulary as opposed to general vocabulary in CLIL settings. This study is therefore imperative due to the dearth of research available about CLIL and academic vocabulary.

The first study conducted by Várkuti (2010) compared CLIL and non-CLIL secondary school students in Hungary and tested their BICS and CALP proficiencies. CLIL students were shown to have significantly higher (by 23%) English competence than their non-CLIL counterpart in CALP-relevant tests. They also showed to have a larger vocabulary size and more sophisticated vocabulary knowledge. In the second study, Agustín-Llach (2015) conducted a study targeting Spanish young (4th grade primary) learners and their vocabulary production in CLIL and traditional EFL settings. The lexical profile analysis revealed CLIL group to have a higher percentage of academic vocabulary usage than non-CLIL in their writings, hence adding up to empirical evidence that is in favor of CLIL.

Lastly, Xanthou (2010) compared learners' content-specific vocabulary knowledge amongst three groups of homogenous 11-year-old learners: a control group who were taught geography through L1 Greek, an experimental group which was an EFL class where students were given L1 equivalents of the target words, and the second experimental group where CLIL approach was adopted. The pre-post-tests measuring students' receptive content word knowledge showed that the CLIL group outperformed the other two groups although all three groups showed significant improvement in the post-test scores. It is noteworthy that unlike many other CLIL

studies, the total exposure time to target words was kept equivalent between CLIL and the EFL groups. Besides, the subjects were selected and grouped in random manner to reduce sampling error. These thoughtful methodological considerations enabled the results to be credited to the positive evidence of CLIL.

Looking at the positive outcomes from the studies above, CLIL might seem to be an ideal alternative to traditional language learning. However, each pro-CLIL argument comes with a counter-argument in other studies (Bruton, 2011). It is therefore risky to attribute all the aforementioned positive results to CLIL exclusively without considering its flip side.

2.4.2 Contradictory Results against CLIL

Overall, positive results of CLIL outweigh its negative evidence. However, there are also studies that undercut the novelty of CLIL which are worth-mentioning. This section will begin by reviewing some of the studies that showed no significant effect of CLIL on vocabulary learning, insinuating that the supremacy of CLIL-effect might have been overstated.

Even though the forenamed Agustín-Llach's (2015) study of CLIL Spanish 4th graders showed a higher academic vocabulary production than their non-CLIL counterparts in their writings, when it comes to receptive vocabulary size and general productive vocabulary, CLIL group did not perform significantly better than the non-CLIL group despite the fact that CLIL learners received nearly 300 hours more English language exposure than traditional learners. This finding seemed to contradict with numerous CLIL studies and showed that CLIL does not always warrant better results than traditional EFL.

Apart from that, Arribas (2016) conducted a receptive vocabulary test amongst CLIL and non-CLIL Spanish students, ranging from the first to 4th year in secondary

school. The author admitted that "the main difference between CLIL and non-CLIL groups lies in the fact that the former received additional hours of exposure to English language by means of CLIL methodology… (and) the number of hours is far from being equal…" (p.278-279).The result showed that both groups of students' receptive vocabulary tests reported no significance difference, even though in general CLIL group received many more hours of English language instruction.

Besides, Olsson (2015) carried out a comparative study of English academic vocabulary use between CLIL and non-CLIL upper secondary school students in Sweden and measured students' productive academic vocabulary using Academic Word List (AWL; Coxhead, 2000) and Academic Vocabulary List (AVL; Gardner & Davies, 2014). To control for the initial differences between two groups, a pre-test was given and it unveiled that CLIL students were already using advanced academic vocabulary compared to the non-CLIL group. Apart from that, over the course of three years, CLIL students did not show more improvement in their usage of academic vocabulary; thus, the CLIL effect was not significant despite having their initial exhibit of promising academic vocabulary knowledge and extra English language exposure in school. Therefore, it tampers the common belief that CLIL is the best-fit methodology and it serves as a warning to wholesale adoption of CLIL (Bruton, 2013).

Lastly, Tragant, Marsol, Serrano, and Llanes (2016) addressed the cross-group comparability issue that many comparative studies in CLIL tend to overlook such as number of instruction hours and characteristics of students that might result in favoritism towards one group over another. To curb this incomparability issue, Tragant et. al. (2016) focused on a single group of young children who went through both CLIL instruction and non-CLIL traditional EFL instruction to compare their productive vocabulary knowledge in two different settings. This method is praise-

worthy for strengthening robustness of the study. The result showed that while the students demonstrated significant growth of vocabulary in both settings, they in actual fact learned more words from EFL lesson than from the CLIL lessons. Therefore, it appeared that CLIL context had failed to provide a better learning environment for students to learn vocabulary than the traditional EFL method. Nevertheless, it is noteworthy that the difficulty levels of the target words that students were given in CLIL and EFL were not comparable such that the analysis with Lextutor program revealed that CLIL textbook contained more challenging academic vocabulary and abstract nouns that EFL material. This might be able to explain the disparity of students' productive vocabulary results in CLIL and EFL settings.

2.5 Call for Critical Assessment of CLIL

While many studies show positive correlation between CLIL and learning of L2/FL, it is also an unmissable reality of the polarized viewpoints on the benefits of CLIL. A possible way to reconcile the seemingly contradictory research outcomes is to critically assess the evidence in the past empirical studies and the interpretations of the results. Bruton (2011, 2013) boldly claims that many empirical CLIL research have failed to justify the benefits of CLIL, and believes that it is the many other factors rather than CLIL alone that contribute to those positive language learning outcomes. As a result, the conclusions might appear to be dubious due to the overlook of those influential factors which then translate into methodological issues. It is proven by a significant number of studies that there are other variables which need to be accounted for when it comes to making any comparison between CLIL and non-CLIL instruction in order to determine whether it is just CLIL or other factors that come into play when CLIL-effect is concerned. While there might be many more other factors that affect

language learning, for this present study, three prominent issues will be discussed and taken into serious account in the methodological design to fortify validity of this study.

2.5.1   Selection Effect

In most countries (e.g. Germany, Finland, Spain, etc.) CLIL is typically optional and the enrollment happens on a voluntary basis. Even though CLIL claims to be open to all learners, many studies show that branding of CLIL as an innovative and effective language learning approach usually attracts learners who possess certain traits, such as students who are already keen in learning the target language (Lasagabaster, 2008). Those interested students and their parents either apply to schools with a CLIL-program or opt to be in CLIL classes for certain subject areas. However, voluntary participation does not always guarantee a spot in CLIL intake because when there is a surplus of applications or due to other logistical predicaments, schools usually select CLIL candidates based on their perception of likeliness to succeed in CLIL although this kind of explicit selection is not necessarily sanctioned by the education authority. Therefore, whether the selection process is implicit or explicit, the mere fact that CLIL participation is not mandatory to all has unavoidably favored students who come with better prior academic achievement, higher motivation in learning, higher cognitive abilities, more favorable family background, etc. (Bruton 2011, 2013; Dallinger et al., 2016a; Sylvén, 2014; Tragant et al., 2016). All of these traits are important prerequisites for learning (Wang, Haertel, & Walberg, 1993 in Dallinger et. al., 2016a), giving CLIL students an advantage over their non-CLIL peers (who are presumably less proficient or less motivated in learning a target language), contributing to a self-fulling prophesy (Dalton-Puffer & Smit, 2013) by mistaking selection effect as CLIL-effect.

The selection bias of CLIL has already been confirmed empirically. For instance, in Sylvén's (2014) baselining study on CLIL and non-CLIL students' proficiency, when comparing the CLIL and non-CLIL students at the start, it was revealed that CLIL students had already outperformed their non-CLIL peers in the vocabulary test. This is because attending a CLIL class in Swedish context is a voluntary choice so while students who find English difficult would not choose to attend CLIL classes, those who choose to participate in CLIL are those who are likely to learn English better. Therefore, it verifies the existence of selection effect of a CLIL program i.e. its tendency to attract students who are already proficiency in the target language. Apart from Sylvén's (2014) study, Várkuti (2010) also admitted that in her study, CLIL students possessed higher language proficiency than non-CLIL group, hence putting both groups on an unequal ground. Therefore, to overcome the comparability issue caused by CLIL/non-CLIL selection, Várkuti had to equalize both groups according to factors that might contribute to target language learning: students' attitude toward learning, general academic achievement, achievement in English class, attitudes toward English learning etc. [See Alonso, Grisaleña, & Campo (2008) for onset inequivalent CLIL/non-CLIL comparison].

While researchers like Sylvén's (2013) and Várkuti (2010) were aware of selection effect and were critical when making conclusion about CLIL, Moghadam and Fatemipour (2014) and Merikivi and Pietilä (2014) did not seem to address the selection issue even though it was highly relevant to their studies. First, in Moghadam and Fatemipour's (2014) study, the researchers chose to compare CLIL students from a private school with a group of non-CLIL students from a public school. They concluded that CLIL method worked because the CLIL students scored better in the vocabulary test. Even though there was no pre-test to evince any disparity CLIL and

non-CLIL groups in terms of cognitive abilities, language proficiency etc., it was highly possible that selection effect existed since CLIL students were from a private school while non-CLIL group was from a public school. It can be assumed that students from the private schools are from higher income families and parents of high socioeconomic status are usually more involved in their children's learning (Roff, 2006). Therefore, this study appears to be over-simplistic by attributing all the language gain to CLIL method alone. As for Merikivi and Pietilä's (2014) study, the subjects consisted of CLIL and mainstream students from the same school. The mechanism of students' assignment to either CLIL or non-CLIL programs was not mentioned but certainly some form of selection had to take place; it was either on a voluntary basis or through the entry exams. Therefore, the non-CLIL students were most possibly the "remnants" in the school who either did not meet the minimum requirement to be eligible for CLIL or those who did not choose CLIL due to fear or lack of interest. Hence, it was highly possible that the prominence of vocabulary learning outcome exhibited by those CLIL students was enhanced by the cognitive and linguistic advantage CLIL students already possessed prior to the program [see Agustín-Llach (2015) with similar situation].

Lastly, De Diezmas (2016) conducted a study in Castilla-La Mancha, Spain that investigated the benefits of CLIL in developing language competences (reading, writing, listening and spoken production and interaction) as opposed to mainstream EFL lessons. However, the findings were not in favor of CLIL as CLIL students did not surpass their non-CLIL peers as expected in many areas. Interestingly the author ascribed this disappointing outcome to the absence of any selection process for the CLIL learners in her study. Unlike most of the places, Castilla-La Mancha had a unique context such that selection of students was prohibited and all students in European

Sections of primary schools were mandatorily enrolled in the CLIL program. Contrary to De Diezmas's view, the researcher perceives this lack of outperformance of CLIL students as a credible evidence against CLIL given that the selection effect is absent in this study. It also further exemplifies the need to control for selection effect when making comparison between CLIL and non-CLIL, for fear that the selection effect is mistaken as CLIL effect.

2.5.2 Unequal Instruction Hours in School

It is generally agreeable that there is a positive correlation between vocabulary development and the amount of exposure language learners receive either in or outside the school hours. A significant amount of research on vocabulary instruction (e.g. Schmitt 2000; Rott, 1999) show the incremental nature of ESL/EFL learners' vocabulary size which grows proportionally to the amount target language exposure and the hours of instruction. For instance, in Schmitt's (2008) review of research studies on post-secondary EFL learners, their vocabulary size ranged from 1,200 to 2,000 word families depending on the number of hours EFL instruction students had had received, ranging from 800 to 1,500 hours. Putting the types of instruction aside, it should therefore be unsurprising that when given more L2/FL instruction hours, students' vocabulary size increases because the amount of formal instruction received dictates the amount of vocabulary learned (Webb and Chang, 2012).

Specifically in CLIL contexts, it is not difficult to notice that many of the CLIL students usually continue with their regular English language lessons alongside their other subjects that are taught in English. This is because to compare CLIL and non-CLIL groups, most of the time it is logistically inconvenient to regroup students with the intention to equalize the hours of instruction in both groups [e.g. Arribas (2016) and De Diezmas (2015)]. Therefore at the onset of the studies, because of the content

lessons taught in the target language, CLIL students already have a time advantage over their peers. Unless an experimental design (Xanthou, 2011) is made possible, it is difficult to control for the total hours of instruction in L2/FL in most of the CLIL/non-CLIL comparative studies, so it is often not brought into the discussion. This obvious incomparable hours of instructions thus makes any purported benefits about CLIL less convincing.

Pladevall-Ballester and Vallbona's (2016) CLIL study on receptive skills addressed the instruction hour comparability to ensure a more rigorous study. To guarantee the comparability between CLIL and EFL groups, they strictly controlled the amount of exposure to the target language in all four testing times. When the exposure time between both groups was kept equal, the possibility of CLIL students having the time advantage could be ruled out. The result showed that, contrary to the common belief, CLIL showed no significant difference from EFL group in reading test, and it was even more shocking that EFL group significantly outperformed the CLIL group in listening. This calls into question the validity of the past studies that are quick to affirm the success of CLIL without consider the difference in instruction hours between CLIL and non-CLIL.

Taking Merikivi and Pietilä's (2014) again as the example, the regular mainstream subjects received much less hours of instruction in English (330 and 600 hours) than the CLIL subjects (2600 and 3400 hours) and the disparity is conspicuous. It is therefore almost foreseeable that the scores of CLIL students would be significantly higher than the regular mainstream group. The authors therefore attributed the high score of CLIL students to the higher exposure to English in CLIL setting compared to the regular learning environment. Due to the seemingly high correlation between hours of instruction in target language and language performance,

for this reason, future research should acknowledge the number of hours of instruction as one of the variables that shape vocabulary learning; otherwise, any positive evidence from the research might not carry as much weight because one can easily argue that the favorable outcomes are actually caused by having the extra instruction hours, rather than the type of instruction (CLIL/non-CLIL).

Agustín-Llach and Alonso's (2016) longitudinal study measured receptive vocabulary size and growth of CLIL and traditional EFL learners for 3 consecutive years, starting from 4th grade in primary school to 6th grade. It disclosed that both groups learned English as a school subject since the first grade of primary school (105–110 hours per year) but on top of that, the CLIL group had been receiving extra-exposure to English through science for an additional two hours a week since their 1st grade (72–74 more hours more yearly) as well. Thus, by the time the study started, the accumulated hours of exposure to English for both groups differed vastly by approximately 300 hours. Results showed that CLIL learners possessed significantly higher overall receptive vocabulary sizes even though the rates of growth in both groups were almost similar. Albeit CLIL seemed to favor receptive vocabulary growth, the authors acknowledged that:

> "There might be other factors influencing students' receptive vocabulary acquisition, such as for example exposure time. It is indeed extremely difficult to discern whether the CLIL factor or the longer exposure (more instruction hours) is to be made responsible for the CLIL advantage in our data, as these two factors are impossible to untangle in our present data" (p.11).

Therefore, this supports the motion to consider instruction hours as a considerable contextual variable in any CLIL research.

On the other hand, even though De Diezmas's (2015) CLIL experimental group received 250 hours in total extra exposure to English in school than the traditional EFL group, its results were not as impressive as expected. The author speculated that it was:

"…due to the number of hours of CLIL instruction received by the CLIL group (only 250 hours in four years), which could have been insufficient for all the language competences to be developed to a significant degree and bring about all the positive effects connected to the use of CLIL methodology" (pp. 97-97).

While it might be true that for CLIL to take effect it needs more time for the results to be more observable, but in order for the CLIL-effect to be justified, the instruction hours should always be kept the same to ensure no time advantage for any groups. This is because the extra instruction hours made available through CLIL should not be seen as CLIL-only advantage [as it implied by Pladevall-Ballester & Vallbona (2006) and Catalán & Ruiz de Zarobe (2009)] because increasing the number of English lessons could be done through traditional ESL/EFL instruction as well. Regarding this point, Dallinger et al. (2016a) raised a thought-provoking question on why one should introduce CLIL at all when one could simply raise the amount of regular lessons to obtain the same results. Having said that, it is not the intent of this study to negate all the findings of the past studies that did not take instruction hours into consideration during data analysis and interpretation, nonetheless, admittedly there is a risk for overselling CLIL if the time bonus is not addressed. CLIL learners may have understood more words because they have received English instruction much longer than their non-CLIL peers, giving them more time to accumulate lexical knowledge.

2.5.3 Extramural Exposure to English

Apart from teaching and learning of English during school hours, a few studies have considered learners' use of English outside school when evaluating the effect of CLIL instruction. Along with other studies [Pearson (2004); Sylvén (2004); Piirainen-Marsh and Tainio (2009)] Sundqvist's (2009) postulates in her doctoral dissertation that the amount of exposure to the TL outside of school correlates with students' language skills. Using questionnaire and language diary of 9th graders from three schools in Sweden spanning one year, she investigated the relationship between students'

extracurricular English-related activities and their oral proficiency and vocabulary tests results. With the result that showed an average of 18.4 hours per week spent on English outside of school, there appeared to be a fairly strong correlation ($r = 0.357$) between extra-curricular English and vocabulary. Hence, Sundqvist concluded that spending time on extramural activities in English had a positive effect on students' language competence, especially in oral proficiency and vocabulary, and regarded extramural English as "a pathway to progress in English" (p.75).

In her thesis, Sundqvist introduced an important term- 'extramural English' to describe in essence, any type of situation in which the English that learners come in contact with outside of the classroom. There are many ways to explain how extramural English can help with vocabulary acquisition, but the most commonly seen is coming from the incidental vocabulary acquisition perspective where it is believed that leaners "pick up" new words when they are engaged in reading, listening, speaking or writing tasks, without conscious intention to commit a word to memory (Rott, 2012) and it occurs in small increments through repeated exposure to words in meaningful context over an extended period of time (Taboada & Rutherford, 2011). Out of the four skills, many research (e.g. Nagy, Anderson, & Herman, 1987; Krashen & Brown, 2007) seem to emphasize on learners' reading-related experiences for incidental vocabulary learning to take place and it is believed that vocabulary growth is determined by the volume and frequency of the child's reading. In other words, incidental learning may occur as a byproduct of any language-related cognitive activity, but most of time it happens through reading (Rott, 1999 & 2012). Besides, it is worth-pointing out that extramural English does not necessarily involve merely learners' undeliberate intention to acquire English (e.g. reading a novel for leisure); it can happen with intentional effort to improve their English as well (e.g. attending tuition classes)

(Sundqvist, 2009). Therefore, as far as extramural exposure is concerned in this current study, both intentional and incidental learning opportunities are taken into account because whether is a conscious or subconscious process, it still impacts L2 word acquisition; it is a threat to the validity of claims made about CLIL.

When it comes to research that look into extramural English exposure in CLIL, one of the major studies is Sylvén's (2004) study that showed Swedish CLIL students' stronger out-of-class reading habit than non-CLIL group when she first investigated CLIL and non-CLIL learners' lexical proficiency. It was shown that those who were involved in out-of-class reading scored higher on vocabulary tests than those who did not read. That was deemed as the possible explanation as of why CLIL students in Sweden had possessed larger vocabulary size than their non-CLIL counterparts since the outset of the study, contributing later to better results by CLIL than the control group. Therefore, Sylvén was hesitant to attribute the success exclusively to CLIL because the additional exposure to English also had a role to play in CLIL group's good performance. This study is significant in pointing out two things: first, it was shown that extramural exposure to English can be one of the most influential factors on vocabulary acquisition because reading habits stand as a stronger predictor than mere CLIL participation (Dalton-Puffer, 2008); second, the fact that CLIL group had already shown a stronger tendency to read outside of school prior to CLIL program, it strongly indicates that extra exposure to English is not resulted from CLIL-instruction as some researchers might claim [e.g. Pietilä & Merikivi, 2014)].

A series of studies then unfolded regarding CLIL students and extramural English; Sylvén (2006) (as cited in Juan-Garau, Prieto-Arranz, & Salazar-Noguera, 2015) reaffirmed the claim that CLIL students were substantially more exposed to English outside school because they read English books and checked web-based

material in English twice as much as non CLIL students. The impact of free-time reading on English vocabulary development has been further confirmed by Finnish researchers, Pietilä and Merikivi (2014) who reported similar observation about how the frequency of reading activity significantly affected students' receptive and productive vocabulary knowledge and that their CLIL participants read English outside school to a larger extent than their mainstream peers. Looking at these evidence, it can be said with certainty that the exposure to English outside school is a variable that needs to be taken into account in any CLIL research to ensure the upper hand one has due to the extra exposure outside class can be minimized because what the learners do in their spare time is germane to their vocabulary knowledge.

2.6   Conclusion

To sum up, this chapter begins by expounding 'CLIL-effect' supported especially by the natural learning approach and second language acquisition theories. Then, it moves on talking about academic vocabulary as one of the many dimensions that make up a language under the CALP/BICS framework. Lastly, a large part of this chapter is dedicated to reviewing some of the significant past studies on CLIL and vocabulary learning. Several factors were not given adequate consideration in the past when making hefty claims about CLIL effect. It is therefore suspected that some of the CLIL-positive results in the past could have been skewed by selection bias, extra instruction hours in school and extra exposure to English outside of school. Overlooking these three factors in any CLIL study may cause a comparability problem between CLIL and non-CLIL groups. Hence, in this study, extramural exposure to English, selection effect and instruction hours will be controlled for and carefully considered before any verdict is made about CLIL and language learning, especially pertaining to academic vocabulary acquisition.

CHAPTER 3

METHODOLOGY

3.1   Introduction

The present study focuses on learners' academic vocabulary learning in CLIL environment by comparing the academic vocabulary performance of three groups of Malaysian secondary school students who either enrolled in the local CLIL program namely DLP or the mainstream curriculum. Simply put, the independent variable in this study is the type of learning environment (CLIL or non-CLIL); while the dependent variable is students' academic vocabulary test performance. Nevertheless, the interpretation of CLIL effect is not as straightforward as just comparing the two groups. The matter is complicated by, as the preceding chapter has discussed in depth, the three potential interfering variables (selection effect, inequivalent instruction hours and extramural exposure) that are shown to be important predictors in any language gain but have often been overlooked in many past studies, hence increasing the risk of the misconstrue of CLIL-effect and its uniqueness. Pérez-Cañado (2012) encapsulates this situation well−there is a "need of solid empirical research which builds in rigorous assessment of the variables under scrutiny […] to determine whether the gains observed are truly ascribable to CLIL practice" (p. 330).

Therefore, it is of this study's primary concern to minimize the intervening effect of these three variables before making verdicts regarding CLIL effect from the comparison of academic vocabulary performance between CLIL and non-CLIL groups. Many decisions in this methodology chapter, from the research design to instrumentation, pivot on the positivist research paradigm and quantitative approach this study has adopted. Therefore, it is important for this chapter to begin by discussing some of the assumptions that are associated with the positivist paradigm before moving

on to justifying research design, selection of subjects and instruments, and statistical tests to answer each research question. Besides, data collection procedures, ethical consideration, pilot study, and validity and reliability issues will also be discussed, and this chapter will end with subject portfolio.

3.2   Research Design

Looking at the numerous empirical research on CLIL, in general, there are more quantitative studies than qualitative studies. However, as it has been discussed thoroughly in Chapter Two, some of these quantitative studies presented weighty methodological flaws. Therefore, it is the intent of this study to contribute to CLIL community by conducting a cross-sectional mixed-method study that investigates critically the effect of CLIL on academic vocabulary at the same time carefully addressing those methodological issues.

One of the reasons for CLIL scholarship to be so diversified is because of the different paradigms or "worldviews" that a researcher adopts and each paradigm comes with its own assumptions regarding "what is reality?", "how do you know something?", and "how do you go about finding it out?" In essence, a research paradigm gives meaning and necessitates certain steps to achieve the goals of study. Therefore, it is paramount to specify the adopted perspectives of a researcher (where the researcher is coming from) in order to facilitate understanding within the research community.

A study that is predominantly quantitative like this is most suitable to be associated with positivist paradigm and its assumptions. From the lens of positivism, it is believed that there is a single reality which is objective, value-free and 'out there' unaffected by the researcher, hence can be measured and known (Creswell, 1994). This is the reason why it is more likely for positivist paradigm to use quantitative methods

to measure and explain this reality with a central goal to draw generalizations from a phenomenon leading to prediction by testing out a theory composed of variables measured with numbers, and analyzed with statistical procedures (Mackey & Gass, 2005; Creswell, 1994). Therefore, validity and reliability are extremely important to enhance the ability to generalize about the issue under study.

Consistent with the assumptions of a positivist quantitative paradigm mentioned above, this cross-sectional comparative study compared DLP and non-DLP groups with two quantitative data collection methods. It was cross-sectional and non-experimental in nature because of the time constraint stipulated by the writer's Master candidature and inconvenience to disrupt the existing class arrangement in normal school days. According to Mackey and Gass (2005), not all studies require a control group and an experimental group or necessitate a pretest and a posttest. However, the validity of a one-shot design like this could be enhanced by having well-identified variables.

This study primarily employed a quantitative data collection method due to the nature of the problem such that it was derived from the existing body of literature with known variables and theories, on top of other logistical considerations including time, manpower and resources. With that said, however, qualitative research method was also equally necessary for triangulation purpose to strengthen the findings from the said quantitative approach with more holistic and descriptive data in order to arrive at a more nuanced conclusion.

3.3   Population of the Study: DLP in Selangor

Dual Language Program (DLP) started with about 296 schools in the first cohort in 2016; however the number has increased exponentially 3 years after its launch. The latest data available tabulated a total of 1429 DLP schools nationwide as of 2018

(Zulkefli, 2018). In Selangor state alone, the number of secondary DLP schools increased from 14 schools in 2016 to 67 schools in 2018. Selangor DLP schools were chosen to be the population of this study because Selangor has the most number of DLP schools amongst all states, and because of its proximity to the headquarter of Ministry of Education, Selangor DLP schools should be more closely guided and frequented by DLP coordinators from MOE for "Program Pemantauan dan Kunjung Bantu DLP" (monitoring and support program to supervise and coordinate DLP implementation).

As it is important to provide a clear depiction of the geographical and cultural context of this study, the following paragraphs are dedicated to provide some background information about Selangor state. First of all, Selangor is one of the 13 states of Malaysia, bordered by the capital of Malaysia, Kuala Lumpur and also less modernized states like Perak, Pahang and Negeri Sembilan. It is the most developed state in Malaysia with the largest population (estimated 6.3 million as of 2016), largest economy but also the lowest poverty rate ("Selangor", n.d.) in the country.

The economic activities and the development of district are often closely linked. The major contributors of economy in Selangor are commerce, industry and services that are well-dispersed across the state. Some of the industrial areas in Selangor are Kajang, Rawang and Selayang, just to name a few. Alongside industries, agriculture is a thriving sector of Selangor's economy, for instance, the cultivation of paddy in districts like Kuala Selangor and Sabak Bernam. Amongst the 67 DLP secondary schools in Selangor, there is a good mix of urban schools (in major cities like Shah Alam, Klang and Petaling Jaya) and suburban schools (e.g. Sabak Bernam and Sepang).

Nevertheless, modernization and globalization in Selangor is ubiquitous with its ever-increasing infrastructure. For instance, to date there are more than 50 malls pervasively everywhere in Selangor and many more are still under construction. Besides, there are 4 public universities and close to 30 private universities and colleges in Selangor alone. Mobility in the state is also tremendously improved over the years as a result of upgraded public transport system, accessibility of highways and e-hailing services such as Grab. Therefore, taking all these together, the differences between urban and sub-urban schools in Selangor might not be as vast as it would be in other less developed states due to the fact that Selangor has collectively higher Human Development Index (HDI) or higher social and economic status than many other states.

This study involved only secondary schoolers. From the previous chapter it was mentioned that unlike the everyday conversational language, academic language involves higher level of cognitive activities and it is often embedded in abstract or context-reduced content. For this reason, the selection of secondary school students instead of primary students is more appropriate. In the primary grades, pupils usually learn the basic skills of language (e.g. learning how to follow instructions in the classroom); while the content learning is mostly taught through visual, manipulative, and experiential means (Sylvén 2013). Only as they grow older will they be encountering language instruction that promotes higher levels of conceptual development (such as word problems in Math and write-ups for Science experiments), and requires them to think, talk, read, and write about content using academic language. Some of the CLIL research also show that older learners are at a cognitive advantage, therefore benefitting more than the young learners. For instance, in response to the lack of significant impact from CLIL instruction amongst her primary school learners, De Diezmas (2015) concluded that it might be due to the underdeveloped cognitive

functions of the young children so it might be necessary to wait until they are older to take full advantage of CLIL methodology. Besides, Muñoz (2006) showed that her older group Catalan students demonstrated higher accuracy rates in English than the younger group with the same number of hours of exposure. That being the case, selection of secondary school students over primary school pupils is justified.

3.4   Selection of DLP School and Non-DLP School

3.4.1   Existence of Selection Effect in DLP School

Prior to this study, the researcher had conducted an informal interview with a DLP teacher who was also the program coordinator in school to establish a preliminary understanding about DLP implementation on the ground. When the circular from the ministry of education first arrived in 2016, the school conducted a survey amongst parents to gauge their interest in letting their children participate in DLP and received an overwhelming number of application. Unfortunately, the school could only allocate one DLP class so the school had no choice, but to be selective of the participants for those 40 spots. In order to be qualified for DLP class, candidates were given tests in English, Math and Science, and only the top scorers were selected. While it was acknowledged that it was the students' freedom to choose to learn Math and Science in English, the coordinator expressed her regret for not being able to reconcile the high demand from the parents with the limited capacity and resources available for DLP. It is an extremely important piece of information because it attests to the selection bias mentioned in the literature review that not every interested student gets the opportunity to be enrolled in the program. Hence, it can safely be assumed that Malaysian CLIL students are either more academically gifted or are already performing better in English, Math and Science subjects than the rest of their non-CLIL peers in the same

school. Therefore, comparing CLIL students with their same-aged, non-CLIL counterparts in the same school is bound to cause comparability issue.

3.4.2 Rationales behind Selection of DLP/Non-DLP Pair

Therefore, when the first research question asks, "Is there any significant difference in the academic vocabulary performance between DLP and non-DLP groups given that the selection effect is controlled?", considering both groups are of the same age, to minimize the selection effect that might have resulted from class streaming in DLP school, the pair must come from two different schools [e.g. School A (DLP) and School B (non-DLP)].

With that in mind, the search for well-matched DLP and non-DLP school-pair began. There are nine districts in Selangor: Gombak, Klang, Kuala Langat, Kuala Selangor, Petaling, Sabak Bernam, Sepang, Hulu Langat and Hulu Selangor. In each district there is a district education office (Pejabat Pelajaran Daerah, PPD) that oversees all matters of the schools within the district. There are a total of 10 PPDs because Petaling Jaya district is split into PPD Petaling Perdana and PPD Petaling Utama.

To begin with, a DLP school was first randomly selected out of the 67 DLP secondary schools in Selangor. However, this research is only be meaningful if: (1) DLP students do not have English as their first language; (2) DLP Math and Science teachers only use English in class. Recognizing the possibility that not every school would meet the criteria set by the researcher, the researcher visited the prospective school to find out more about the school background, especially their DLP students' first language because if there were too many English-speaking students in the DLP group, it would make up a small eligible sample, therefore not ideal for a quantitative study like this. The researcher went through two rounds of DLP school selection

because in the first chosen DLP school, most students regarded English as their first language.

Once that DLP school was confirmed, the researcher then looked for a comparable school that did not provide DLP option. It was done first by screening for non-DLP schools that were located in the same district as the target DLP school; schools that were too far away from the DLP school were filtered out in order to increase the likelihood of DLP and non-DLP schools having comparable socio-economic background between. However, having a non-CLIL school within the same district as CLIL school was insufficient. The next step was to ensure the comparability in terms of students' academic competence. To do so, the researcher sought help from an officer in a relevant department who allowed the researcher to have access to the *Gred Purata Sekolah* (GPS, a common academic performance indicator for Malaysian secondary schools) of all schools in the district. The non-DLP school with the closest GPS to that of DLP school was selected; besides, only the best performing class was involved.

School A (DLP) and School B (non-DLP) were finalized after many layers of control throughout the selection process. Therefore, both DLP and non-DLP groups could at least be assumed equivalent in their socio-economic background, general cognitive abilities and attitudes towards learning. Nonetheless, their prior English proficiency needed to be checked using their UPSR results only after they had answered the questionnaire.

3.4.3 Why Form 2 Students, instead of Form 1 and 3?

At the point of data collection (January 2018), there were three cohorts of secondary DLP students available: the 2016, 2017 and 2018 cohorts. Needless to say, it was too early to study the 3rd cohort when they just started the program; the researcher chose

the second DLP cohort (Form Two in 2018) over the first cohort (Form Three in 2018) because of the new curriculum they were using. Starting from early 2017, the revised Standard Based Curriculum for Secondary Schools (KSSM) and Standard Based Curriculum for Primary Schools (KSSR) were supposed to replace the old curricula KBSM and KBSR. However, when DLP was first implemented in 2016, the new Math and Science textbooks in English were not in-print yet so students had to resort to using both *Malay* KBSM textbooks and obsolete PPSMI textbooks. Hence, the researcher decided to study the second DLP cohort (2017) because the opportunity of the first cohort to pick up academic vocabulary might have been jeopardized by not having a formal textbook for a year.

3.4.4   Why is Form 2 DLP Compared with Form 4 non-DLP?

According to the administrators in both School A and B, conventionally there are 7 English, Science and Mathematics lessons per week for each subject with approximately 30 minutes per lesson, but both Form 2 DLP and non-DLP's English credits to 8 lessons per week under the new curriculum. Even though the exact total number of hours cannot be known, it still does not change the fact that DLP students in general receive many more formal instruction hours in English. Assuming that DLP and non-DLP groups had equal amount of instruction hours during primary school time, with the extra 14 Math and Science lessons (7 hours extra) Form 2 DLP students were receiving each week (assuming 40 weeks of school per year), at the point of data collection in the first week of January 2018, they had had received 280 hours more formal English instruction compared to their same-age non-DLP peers.

Therefore, to answer research question 2, the number of English instruction hours between DLP and non-DLP must be comparable, but it can only be done by comparing Form 2 DLP to Form 4 non-DLP who were two years older (instead of the

same-age peers) because Form 4 non-DLP students would have had accumulated compatible total instruction hours at same data collection time [see Table 3.1]. Selection of subjects from different grades but with similar instruction hours is also seen in Alonso's (2015) study.

Table 3.1

Estimation of Total Instruction Hours in English in Early January 2018

School	Group	Estimated Total Instruction Hours* in English
A	Form 2 DLP	440 (8 English+ 7 Math + 7 Science in Form 1)
B	Form 2 non-DLP	160 (8 English in Form 1)
A	Form 4 non-DLP	420 (7 English from Form 1 to Form 3)

Note. 30 minutes per lesson; 40 weeks per school year.

In Table 3.1 it is obvious that Form 2 DLP and Form 4 non-DLP's total hours are not exactly the same; it is inevitable because it is not possible to precisely track down actual instruction time for each lesson due to many extraneous factors (such as prolonged school assemblies that spilled over into official lesson time and forfeited lessons because of special school events, just to name a few). Therefore, for practicality sake, Form 2 DLP and Form 4 non-DLP were assumed to have comparable formal English exposure in school. Lastly, on a side note, these two groups can be of the same school because there was no Form 4 DLP in that school so this Form 4 non-DLP group was indeed the best class across the board. In fact it is better to have these two groups coming from the same school to minimize the external factors, especially students' socio-economic background.

3.5 Restatement of Research Questions and Subject Groups

This research was guided by three research questions to: (1) compare the results between a DLP school and non-DLP school, having the selection effect controlled for; (2) compare again a DLP group with a non-DLP group when their formal English

instruction hours are kept constant; and (3) compare DLP and non-DLP students' exposure to extramural English in order to investigate if extramural English was a confounding factor. From the sections above, it was made clear how selection effect and formal instruction hours in English were controlled for in order to minimize potential confounding effects. As for the third predictor—extramural exposure, it was controlled by first checking if both DLP and non-DLP groups showed any discrepancies in their English-related activities after school. This was to ensure that neither groups had any extra exposure to English outside of school that might skew the results. It was hoped that these carefully thought-out questions and stringent control for confounding factors would help unmask a more "distilled" CLIL-effect. These questions were answered mainly through two instruments: academic vocabulary test and questionnaire, administered to three (3) groups of students from a DLP school and a non-DLP school. The research plan is summarized in Table 3.2 below:

Table 3.2

Summary of Research Plan

Research Questions	Subjects	Schools
Is there any significant difference in the academic vocabulary performance between **Form Two DLP and Form Two non-DLP** groups given that the **selection effect** is controlled?	Form 2 DLP Form 2 Non-DLP	A B
Is there any significant difference in the academic vocabulary performance between **Form Two DLP and Form Four non-DLP** groups given **their total instruction hours in English** are controlled?	Form 2 DLP Form 4 Non-DLP	A A
Is there any difference between DLP and non-DLP students' **extramural English exposure**? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English a confounding factor in this study?	Form 2 DLP Form 2 Non-DLP Form 4 Non-DLP	A B A

3.6 Data Collection Procedures

The preparation for data collection began officially after the vetting process of the research proposal by the Faculty of Education in University Malaya. In order to carry out any research in government educational institutions, approval from the Education Planning and Research Division (EPRD) or *Bahagian Perancangan dan Penyelidikan Dasar Pendidikan* (BPPDP) of the Ministry of Education (MOE) is compulsory. After gaining the approval from the EPRD [Appendix 1], another approval was sought from the state education department or *Jabatan Pelajaran Negeri* (JPN) [Appendix 2]. Once the permission letter from JPN was received, the researcher met with school principals to seek for their permission to conduct the field studies. The purpose and nature of this study were explained thoroughly and the researcher also took the opportunity to ensure

that the information about the schools were accurate. Throughout the process, there were many unfruitful school visits because of various factors such as mismatched information (non-DLP schools were actually DLP schools but the update was not available online) and miscommunication resulting to collecting data from the wrong class (not the best performing class). Due to the these hiccups, the data collection process took longer than expected and the selection process was repeated until fitting DLP and non-DLP pair was found.

Once the final DLP and non-DLP pair was secured, the researcher made several more visits to meet the teachers-in-charge to help them understand the purpose of this research and its process. The meeting and data collection dates were decided by the schools to minimize unnecessary disruption of the normal lessons and teachers' schedule. On the actual data collection days, researcher entered the agreed classrooms to administer the academic vocabulary test and questionnaire to the students. It was conducted in their own classrooms on a normal school day in order for the environment to be as familiar and neutral as possible. The session began by researcher introducing herself and her area of research. Then students were then informed in English and Malay languages about the privacy and confidentiality of this research, and their right of withdrawal from the study at any time throughout the study. After ensuring all relevant information was relayed and understood, a consent form was given to each student for their parents or guardians to give their consent (Appendix 3).

Then a set of questionnaire and academic vocabulary test was distributed. Before starting, the researcher briefed the students on the instruments and clarified some of the questions from the floor. There was no time limit to complete both the questionnaire and the test; the average time taken was 20 minutes. Before the end of

the data collection session, students and teachers were thanked for their cooperation. The next day, the researcher returned to the schools to collect the consent forms.

The data collected were then cleaned and analyzed. To validate and strengthen the findings of the data, a few months later the researcher set up for an appointment with a DLP teacher from the same DLP school to conduct a semi-structured interview. Prior to the interview, terms of confidentiality were communicated and personal consent was gained. The interview was conducted in an empty classroom. The process took around 30 minutes. Lastly, interview was transcribed as a wrap-up of the entire data collection process.

3.7   Ethical Consideration

Ethical issues related to gathering data from school students are gaining more attention over the years. As this research involved many different stakeholders: students, parents, teachers, school administrators and various departments under the Ministry of Education, ethical considerations such as gaining of access, informed consent, privacy and confidentiality needed to be handled cautiously to ensure all steps were in compliance with the policy stipulated by University of Malaya Research Ethics Committee (UMREC) on human ethical practice in research.

3.7.1   Gaining Access

To gain access to three schools (2 for actual studies and one for pilot study) and the respective classes, the protocols set by the Ministry of Education Malaysia were carefully adhered. First, according to the circular released by Ministry of Education Malaysia (KPM, 2006), every researcher has to submit their proposal (chapter 1 to 3) to Educational Planning and Research Division (EPRD) or *Bahagian Perancangan dan Penyelidikan Dasar Pendidikan* (BPPDP), MOE. The review process took about 3 weeks. Once permission was granted, another permission was requested from

Education Department of Selangor or Jabatan Pendidikan Negeri (JPN) Selangor. Then, with the letters from BPPDP and JPN, the researcher met up with the school principals to seek for their approvals.

3.7.2 Informed Consent

Gaining informed consent from research subjects according to Mackey and Gass (2005) entails: 1. full disclosure about the study, 2. comprehension on the part of the subjects, 3. voluntary participation (p.27). To ensure subjects' comprehension about informed consent, considering their young age, it was verbally explained in a language that was comprehensible and meaningful to their level. Besides, as the subjects were all below 18 years old, parental consent in black and white was also needed. The bilingual consent form included research purpose, procedures, potential risks and benefits, steps to ensure confidentiality and anonymity. The researcher's contact number was also printed on the consent form in case of any inquiries.

Lastly, the participation of the subjects was voluntary and they were not coerced into answering the questionnaire and the vocabulary test. The researcher informed the subjects about their right to withdraw their participation without any penalty. They were also given reasonable time to consider and clarify any doubts they had before the tests. At that time, no subject indicated unwillingness to participate.

3.7.3 Privacy and Confidentiality

During the conversations with school principals, it was communicated as a great concern that the school reputation's reputation would be tarnished if any negative evidence against the school administration was published. To ease their worry, the researcher had provided a letter to reassure the school administrators that identities of schools and individuals would not be disclosed to a third party. On the other hand for the students, as anything related to their academic performance was always a sensitive

issue, they were reassured that their academic vocabulary test result would not be shared with their peers, teachers and parents. To ensure that the subjects' identities were kept anonymous and confidential as much as possible, when analyzing and reporting the data, code numbers instead of names were used to refer to the subjects. The researcher was also selective regarding what to be reported in the writing so that the identifying information that might easily give away the identities of the schools was not included. Finally, all the data will be kept solely with the researcher for one year after the completion of her masters' program.

3.8   Instruments

To answer the three research questions, two main instruments-an academic vocabulary test and a questionnaire were employed. Lastly for triangulation purposes, an interview session was conducted with a DLP Math teacher from the same participating DLP school to provide more depth to the study.

3.8.1   Instrument (1): Academic Vocabulary Test

3.8.1.1 Academic Word List (AWL)

 As previously discussed, academic vocabulary generally refers to vocabulary that occurs more often in academic contexts than in other contexts (Baumann and Graves, 2010). The present study is limited to the analysis of general academic vocabulary that appears across all disciplines. Before talking about the instrument to measure subjects' academic vocabulary performance, it is important to first talk about a commonly cited academic word list compiled by Coxhead (2000), namely the Academic Word List (AWL). AWL is a list of 570 word families extracted from a corpus of 3.5 million running words in academic books and journals of different domains such as the Arts and Humanities, Commerce, Law, and Science. For a word to be included in the AWL, it had to adhere to stringent criteria, making it specific to academic contexts. First the

list does not include the most frequent 2000 words of English (the General Service List). Besides, a word had to appear at least 100 times in the academic corpus and at least 25 times in all of the four sections of the corpus (Coxhead, 2000). The AWL has been widely used in a large number of studies, showing consistent coverage of approximately 10% of the total words in academic texts but only 1.4% of the total words in a fiction collection (Coxhead, 2011) [See Appendix 4 for examples of academic words in AWL].

3.8.1.2 Justification of Selection of AWL rather than AVL

Other than Coxhead's (2000) Academic Word List, another well-known list, Gardner and Davies' (2014) Academic Vocabulary List (AVL), was claimed to be more extensive and fine-tuned than AWL with higher coverage of the most frequent academic words in Corpus of Contemporary American English [COCA] and The British National Corpus [BNC] (Olsson, 2015). Coxhead's AWL was often mostly criticized for its use of word families when determining word frequencies because a word family may contain a large number of members with distinct meanings such as the word family 'react' which contains thirteen members but do not all refer to the same definition (Gardner & Davies 2014). On the other hand, AVL is considered more comprehensive because lemmas were used instead of word families when compiled. After much consideration, despite of its deficiency in terms of quantity (570 words in AWL but 3000 words in AVL), AWL was still chosen over AVL in this study firstly because AWL has been used for a longer time in CLIL-academic vocabulary worldwide studies [e.g. Agustín-Llach (2015) and Olsson (2015)], therefore using AWL is beneficial to enable more international comparisons and discussions about CLIL performance and results across different contexts; moreover, the ultimate goal of this study is to establish comparison between CLIL and non-CLIL groups, not to

assess subjects' overall academic vocabulary capacity so even though AWL contains a much shorter list, it is still useful for comparative purposes.

3.8.1.3 AWL Tested Using Vocabulary Levels Test (VLT)

Many types of vocabulary tests have been used to measure English learners' general vocabulary competence but reliable test for academic vocabulary is scarce. For this study, the academic vocabulary test is adopted from Nation's (2000) Vocabulary Levels Test (VLT) [improved by Schmitt et al. (2001)]. There are several functions and sections in VLT but only one part of VLT was used.

To begin with, VLT is essentially a set of frequency-based tests which is commonly used by researchers and practitioners to gauge learners' vocabulary size. It comprises of words from the 2000 level [i.e. the 2000 most frequent words in English (e.g. *victory* and *develop*)], 3000 level, 5000 level and 10,000 level [i.e. much less frequently occurring words (e.g. *benevolence* and *pacify*)]. All the words in each section were selected so that they would be representative of the words at that frequency level. Apart from the frequency-based tests, VLT also includes a test that is not based on frequency levels, specific for academic vocabulary using words from the Academic Word List. In summary, VLT consists of 5 sections: 4 frequency based tests and 1 AWL-based test. According to Kremmel and Schmitt (in press), the AWL section in VLT can be used as an independent test, hence it is an appropriate instrument for this study.

All sections in VLT come with two versions, Version A and Version B. Each test consists of 30 items formed in 10 clusters using a multiple matching format. Each cluster consists of 6 words in the left column and 3 descriptions on the right column. Test-takers are asked to match three of the six words on the left with the corresponding

meanings given on the right. Below is an example taken from Version A 2000-level test [see Appendix 5 for full AWL test]:

```
1 business
2 clock          _____  part of a house
3 horse          _____  animal with four legs
4 pencil         _____  something used for writing
5 shoe
6 wall
```

When it comes to validity and reliability of VLT, several analyses were conducted by Schmitt et al. (2001): item analysis showed that correct answers did reflect test-takers' understanding of the target words; while Cronbach's alphas for all of the levels and AWL in both versions were high (above 0.90). In addition to that, the test for equivalence between Version A and Version B showed small-scale of differences so these two versions could be used as alternate forms. For this study, to preempt any possible differences in scores caused by the version of tests used, only Version A was used. As cheating during the test might be a threat to internal validity, the researcher prepared two sets of test paper with shuffled sequence of the same questions to prevent students from copying each other's answers.

3.8.1.4 Justification for the Selection of VLT

In a nutshell, beside for its proven validity and reliability, the format of VLT suits a comparative study like this because it captures and allows straight-forward comparison of the very basic and initial stages of test-takers' recognition of the form of the vocabulary (breadth). In addition to that, DLP and non-DLP subjects' knowledge of AWL could be quantified in a way that though does not to measure the size of their overall academic vocabulary knowledge, it enables apple-to-apple comparisons with other vocabulary research in CLIL scholarship using the same instrument; amongst the CLIL-vocabulary studies the researcher has encountered, VLT is the most widely used test to provide a snap-shot for leaners' receptive vocabulary performance.

3.8.2 Instrument (2): Questionnaire

One of the most used tools in social science research is self-completion questionnaire. It is useful to elicit information that is not available from production data alone (Mackey & Gass, 2005) through respondents' self-reporting process. It also enables quantitative data to be collected in a standardized way so that the data are internally consistent and coherent for analysis (Malhotra, 2006). Besides, compared to individual interviews, using questionnaire is a more economical and practical method to reach out to a number of respondents simultaneously especially for a study that has limited resources like this. However, the questionnaire was designed by the researcher, the validity and reliability of this instrument must be established. For this reason, a pilot test was imperative to test-run the questionnaire [see pilot study, p.75].

Each subject was required to complete this questionnaire on paper along with the academic vocabulary test mentioned above. They were encouraged to answer truthfully and were reassured of the confidentiality of their response. This would help them to be at ease without the fear of being judged because of their performance. Besides, considering the time constraint during normal school hours and to avoid fatigue in answering, this questionnaire was designed to be as succinct as possible. The pros and cons of using a self-report and self-designed questionnaire like this will be discussed in the following 'validity and reliability' section.

For this study, the main function of the questionnaire was to elicit information on subjects' frequency and types of extramural activities that were shown in many literature to have positive effect on students' vocabulary learning. Students' extramural exposure is a factor that needs to be considered when making conclusion about CLIL effect because extramural exposure can reap similar effects.

3.8.2.1 Questionnare Overview

The first part of the questionnaire (See Appendix 6) was an introduction that explained the aim of this research followed by a set of simple instructions. The questionnaire was composed of mostly closed-item questions that rendered a greater uniformity of measurement and therefore greater reliability. A few open-ended items were also included to allow respondents to express their own thoughts and ideas which then might contribute to more insightful data. The questions were worded using simple and ordinary words that were compatible with the vocabulary level of the respondents. Questions were also arranged in a logical order, from general and nonthreatening questions to more specific and complex questions.

The first part of the two-fold questionnaire asked for basic biodata of the respondents in order to ascertain the right students were answering the questionnaire in terms of their home language and to also establish the comparability of two groups based on their previous academic performance and English proficiency.

The second part, the main focus of this questionnaire, was designed to obtain information about respondents' extramural activities that could potentially be quantified as their extra exposure to English in answering research question 3. The participants were asked to specify how often they were involved in different activities in English, using a consistent frame of reference for each respondent, whether it was *every day, once or a few times a week, once or a few times a month, a few times a year,* or *never or almost never.* These options were carefully thought-out in order to be unambiguous. Subjects were also instructed to choose only one answer for each question. Questions with multiple-choice format like this was easier to analyze and tabulate than open-ended questions. However, there were a few open-ended short

questions for the subjects to list down some examples as supplement to some multiple-choice questions.

3.8.3 Instrument (3): Interview

Lastly, in order to triangulate the data from the aforementioned instruments and to answer the research questions in a more comprehensive manner by adding the perspective from one of the important stakeholders on grassroots level, the researcher conducted a semi-structured interview with a Mathematics DLP teacher, Cikgu Maryam (a pseudonym) to gain more insight into the DLP classroom. Cikgu Maryam was a young teacher with seven years of teaching experience. She was trained as a Math teacher in English medium and interestingly she experienced PPSMI when she was a student. At the point of writing, she had taught DLP for about one and half years.

A list of questions were prepared beforehand to make sure that all relevant topics are covered [see interview protocol in Appendix 7]. However, the researcher was spontaneous in formulating new or follow-up questions that arose during the discussion when Cikgu Maryam disclosed more useful new information data. The entire interview was audio-recorded and then transcribed [sample see Appendix 8].

The researcher chose to conduct the interview with a teacher even though undeniably there were many other stakeholders from high-level policy makers to actual classroom members; as it was evidenced in many past studies (Walker & Tedick, 2000; Tan, 2011; Lyster & Ballinger, 2011), a teacher's actions and decisions in a CLIL classroom were contingent on one's beliefs and understanding about DLP actions. Teachers are the ones who directly interact with students in integrated content and language settings so they play a paramount role in making or breaking a CLIL program. Therefore an interview as such could provide more accurate contextualization required in answering the research questions.

3.9   Validity and Reliability

Internal validity measures to what extent the differences found for the dependent variable (academic vocabulary test result) are directly related to the independent variable (types of instruction: DLP or non-DLP). Therefore, there was a need to control for or rule out all other possible factors that could potentially account for the observed results. Hence, this study was carefully designed with the intention to curtail threats to internal validity as informed by the past literature such as selection effect that favors participants of certain characteristics, and exposure to English in school and outside of school. As for the external validity which concerns the generalizability of the findings to the DLP population in other states that share similar context with Selangor, random sampling was employed from a total of 67 DLP secondary schools in Selangor.

There were two main instruments in this study: the AWL section in Vocabulary Levels Test (VLT) and researcher's self-designed questionnaire. The validity and reliability of the VLT were already well-established by previous studies such as Schmitt, et al. (2001) [see 'AWL Tested Using Vocabulary Levels Test (VLT', p.69]. However, the AWL test had to be pilot-tested to avoid ceiling or floor effects such that the test was either too easy or too hard for the target subjects. The result of the pilot test will be discussed in the immediate section.

As for the questionnaire, the researcher had to go through extra procedures to establish its validity and reliability. There are a few potential shortcomings of a self-report questionnaire. For instance, when it comes to comprehending a question, it is always possible for respondents to interpret the questions differently from its intended purpose or to perceive things from a different perspective. Therefore, for a questionnaire to be effective, all questions have to clear and unambiguous. In order to do so, the researcher asked two peer researchers to review the questionnaire and made

some amendments based on their feedbacks. Apart from that, as the subjects were ESL learners, any English-only questionnaire might present some extent of constraint to their understanding towards the questions. Therefore, to overcome this, a *Malay* translation was provided when students indicated difficulty in comprehending the questions; they were also allowed to answer in *Malay* for the short-ended questions so that language barrier would not impede their ability to disclose information.

3.10   Pilot Test

Pilot test is essential in making sure the instruments are valid and reliable in measuring target variables. It took place in another DLP school and involved 20 best performing DLP students who were of the same age as the actual DLP subjects. The number of respondents in the pilot test was recommended by Malhotra (2006) for pretesting questionnaire. The pretest was conducted in a classroom during the normal school day. The researcher was not permitted by the school to enter the classroom to interact with the students so the instruments were given to a teacher instead. The teacher was briefed by the researcher beforehand and was given a script to read to students. Each subject was given a vocabulary test, questionnaire and consent form. With the intention to improve the questionnaire, the researcher asked the teacher to collect students' feedbacks about the questionnaire verbally or in writing regarding the wording, layout, comprehensibility of questions and clarity of instructions. Students were also asked to underline, or circle words and phrases that they found confusing and ambiguous. As for the academic vocabulary test, each subject was given either version A or B of AWL in VLT in order for the researcher to choose a more grade-appropriate version for the actual study.

The results from these pilot study subjects' vocabulary test questionnaire were then sorted and analyzed. Subjects without the consent form or who had English as

their first language were not included in the sample. The analyses showed that both versions of AWL test in VLT were appropriate for the age group. The mean score was 13.25 (out of 30) with standard deviation of 4.36 so ceiling or floor effect was not of concern. Two groups of subjects who sat for version 1 and version 2 of the test scored extremely closely (mean=13.3 and 13.2 respectively); therefore, there was no preference of which version should be used in actual test.

As for the questionnaire, even though students did not give any verbal feedback according to the teacher on duty, based on the markings and scribbles on the questionnaires, the researcher was able to identify several problematic spots. A few amendments were then made accordingly. Firstly, questions with confusing word choices were rephrased to enhance the clarity of the questions. Secondly, the researcher also realized that there were students' UPSR English came in two papers (comprehension and composition) so sub-questions were added to ask about results for both papers. Besides, extra questions about subjects' primary school name and tuition for STEM subjects were included while a question about "self-initiated research" was removed because based on students' written responses, it was inferred that the question was vague. Lastly, the redrafted questionnaire was shown to two peer researchers to ensure content validity: effectiveness of questions to elicit information about learners' English extramural activities. Then, Person's correlation was performed (without the "self-initiated research" question), given the p-value= 0.058, it was slightly more than the significance level of $\alpha = 0.05$. However, when considered at statistical significance at the 10%, correlation (r=.43) between students' academic vocabulary test result and extramural exposure to English language could serve as an evidence, albeit weak, that extramural exposure had the potential to be a confounding factor so it would be worth controlling for in the actual study.

As for the instrument's consistency, the most appropriate way to establish reliability of the questionnaire was by performing a statistical method using Cronbach's Alpha to assess the inter-item consistency for the measurement of 'extramural exposure to English language' construct. Cronbach's Alpha was chosen over other methods because it can be applied to ordinal data which was the case for this questionnaire as it required respondents to choose from a series of categorical and ordered responses for each multiple-choice question. The reliability test showed that Cronbach alpha was 0.87. According to Hair (2010) quoted in Mohamad (2016), the alpha value of above 0.7 is acceptable, therefore the 14 measurement items in the questionnaire ('self-initiated research' question excluded) were reliable and consistent. Hence, from this pilot test, it could be assured that both instruments functioned well as a whole.

3.11   Procedure for Data Analysis

Before data was analyzed, to ensure that the right subjects were involved in this study, the biodata section of the questionnaire was checked first and foremost. The data of those students who indicated 'English' as the language they spoke at home were excluded. To answer each research question, descriptive and inferential statistics were both be used in analyzing the data. Besides, the data were also checked for normal distribution using Kolmogorov-Smirnov statistical test before either parametric or non-parametric tests could be used depending on whether the sample met the normality assumption.

For research question 1, before independent t-test was employed to compare the means of Form 2 DLP and Form 2 non-DLP's academic vocabulary test scores, comparability between these two groups had to be examined to minimize any interfering effect caused by extraneous factors even though the selection effect had

already been remedied when choosing the DLP/non-DLP school duo. This time, students' genders, types of primary schools, general Ujian Peperiksaan Sekolah Rendah (UPSR) results and UPSR English results were checked. Only then it was safe to assume that the selection effect to be insignificant and both groups were deemed equivalent in terms of their academic giftedness.

Next, for research question 2, independent T-test was used again to compare the means of F2 DLP and F4 non-DLP's academic vocabulary test score. It was to find out if DLP students would still outperform non-DLP without the extra time advantage. Besides, students' extra tuition hours for English and STEM subjects were also checked so that it did not become another intervening factor.

The last research question was two-fold. The first part of the question was answered using one-way analysis of variance (ANOVA) to determine whether there was any significant difference in extramural exposure amidst Form 2 DLP, Form 2 non-DLP and Form 4 non-DLP. Then, Pearson product-moment correlation coefficient or Pearson's r was used to measure of the linear correlation between subjects' academic vocabulary result and the overall extramural exposure. The findings will be presented and discussed in Chapter 4. The data analysis procedure for each research question is summarized in Table 3.3.

Table 3.3

Summary of Data Analysis

Research Questions	Subjects	Statistical tests and steps
Is there any significant difference in the academic vocabulary performance between **Form Two DLP and Form Two non-DLP** groups given that the **selection effect** is controlled?	Form 2 DLP Form 2 Non-DLP	**Preliminary Data Analysis (Steps 1-3)** 1. Independent t-test and one-way Anova to check if there was any significant difference between gender and among types of primary schools. 2. Correlation of UPSR general result and UPSR English result and academic vocabulary test result was performed. 3. When correlation was found significant, two groups were equalized by taking out the outliers. 4. When Form 2 DLP and Form 2 non-DLP's UPSR English results were matched, independent t-test was performed to compare the academic vocabulary test results of both groups.
Is there any significant difference in the academic vocabulary performance between **Form Two DLP and Form Four non-DLP** groups given **the total instruction hours** in English are controlled?	Form 2 DLP Form 4 Non-DLP	**Preliminary Data Analysis (Step 1)** 1. Correlation between students' English and STEM tuitions outside of school and their vocabulary test results was tested. 2. When the correlation was found to be insignificant, independent t-test was done to compare the academic vocabulary test results for both groups.
Is there any difference between DLP and non-DLP students' **extramural English exposure**? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English a confounding factor in this study?	Form 2 DLP Form 2 Non-DLP Form 4 Non-DLP	1. One-way Anova was conducted to check if there was any significant difference between the three groups. 2. Correlation test was done between overall extramural exposure and academic vocabulary test results.

Finally, for the interview data, the researcher listened to the interview recording twice before coding. The interview underwent a two-step coding process (Strauss & Corbin, 1998). First, open coding was done to identify sections containing main themes of interest such as how the teacher viewed oneself as a DLP teacher, how an everyday DLP lesson was like etc. Then, these sections were selectively transcribed. Specific parts of the transcript were coded again in sub-themes.

3.12 Preliminary Data Analysis

3.12.1 Preliminary Data Analysis for Research Question 1

To prepare for the actual data analysis to answer research question 1, potential confounding factors such as subjects' gender, primary school type, and UPSR general and English results needed to be checked. Therefore, a series of inferential tests (independent T-test and Pearson product-moment correlation) were conducted.

3.12.1.1 Controlling for Subjects' Gender and Primary School Type

Table 3.4

Descriptive Data and Independent T-test Result on Academic Vocabulary
Performance based on Students' Gender and Primary School Type

	Group	N	Mean	Standard deviation	t	df	Sig (2-tailed)*
Gender	Male	25	18.68	6.473	1.879	90	0.064
	Female	67	16.16	5.412			
Primary School Type	SK	84	16.77	5.675	-.395	90	0.694
	Boarding	8	17.63	7.328			

* p< 0.05 level of significance

There were 25 males and 67 females in the raw sample. Firstly to check if there was any significant difference between male and female students' academic vocabulary performance, the Independent t-test result in Table 3.4 shows that there was no significant difference between male and female subjects, given t(90)=1.879, p=0.064.

Therefore, the effect of the subjects' gender on their academic vocabulary performance was negligible. As for subjects' primary schools, '*Sekolah Kebangsaan*' Malay-medium national school (SK) and '*Sekolah Berasrama*' (boarding school) were compared and it was also found that there wasn't any significant difference between students who attended national school and boarding school, as determined by t-test shown in Table 3.4, $t(90) = -.395$, $p=0.694$. Hence, primary school type was also ruled out as a confounding factor.

3.12.1.2 Controlling for Subjects' Prior English Proficiency

Even though during the onset of the school selection stage the researcher had tried to match the DLP school with another compatible non-DLP school that was in the same district with similar GPS, when it comes to individual learners' prior academic performance and English proficiency, in the raw data, Form 2 DLP and Form 2 non-DLP were shown to still have discrepancies in the their UPSR overall and UPSR English results to begin with. Based on Table 3.5, Form 2 DLP did better in both UPSR overall (mean=4.46) and UPSR English (mean= 4.5) than Form 2 non-DLP [UPSR overall=3.9; UPSR English=3.76]. To confirm, independent T-test results in the same table showed that $t=4.99$, $p=0.000$ for UPSR overall and $t=4.695$, $p=0.000$ for UPSR English, proving that the differences in UPSR overall and UPSR English were significant in both groups. Therefore, it was evident that the Form 2 DLP and Form 2 non-DLP did not possess similar levels of general academic abilities and English proficiency to begin with.

Table 3.5

Descriptive Data and Independent T-test Result on UPSR Overall and UPSR English Results between Form 2 DLP and Form 2 Non-DLP

UPSR	Group	N	Mean	Standard deviation	t	df	Sig (2-tailed)*
Overall	Form 2 DLP	23	4.46	0.246	4.990	54	0.000
	Form 2 non-DLP	33	3.90	0.559			
English	Form 2 DLP	23	4.5	0.426	4.695	54	0.000
	Form 2 non-DLP	33	3.76	0.751			

* $p< 0.05$ level of significance

Table 3.6

Correlations between Subjects' Vocabulary Performance and UPSR General and English Results

		Vocabulary Score	UPSR General Result	UPSR English Result
Vocabulary Score	Pearson Correlation	1	.658[**]	.743[**]
	Sig. (2-tailed)		.000	.000
UPSR General Result	Pearson Correlation	.658[**]	1	.802[**]
	Sig. (2-tailed)	.000		.000
UPSR English Result	Pearson Correlation	.743[**]	.802[**]	1
	Sig. (2-tailed)	.000	.000	

**Correlation is significant at the 0.05 level (2-tailed)

Table 3.7

Descriptive Data and Independent T-test Result on UPSR English between Form 2 DLP and Form 2 Non-DLP after Matching

Group	N	Mean	Standard deviation	t	df	Sig (2-tailed)*
F2 DLP	22	4.546	0.375	1.848	38	0.072
F2 Non-DLP	18	4.333	0.343			

* p< 0.05 level of significance

Next, Pearson correlation test was conducted to find out if UPSR overall and UPSR English were actually confounding factors since as stated above DLP students were shown to have done better in their overall and English UPSR tests. Results in Table 3.6 below shows that there were actual significant correlations between subjects' academic vocabulary test and UPSR general results [r (92) =.658, p<0.5], and subjects' academic vocabulary test and UPSR English results [r (92) =.743, p,.05], affirming the necessity for students' prior academic results and language proficiency to be controlled for to enable unbiased comparison.

It can be seen in Table 3.6 that subjects' UPSR English result had a stronger correlation (r=0.743) with their academic vocabulary test score than UPSR general result (r=0.658), hence the researcher chose to control for subjects' UPSR English result only (instead of both) to make sure that CLIL effect was not skewed by subjects' prior English proficiency.

To do that, a series of steps were taken to ensure that the DLP and non-DLP groups were matched in terms of their UPSR English results before research question 1 could be answered. Firstly, outliers were removed from the subject sample. Secondly, subjects who scored lower than grade 'B' in their UPSR English were removed so that DLP and non-DLP's UPSR English means could be matched. In Table 3.7, it was shown that after matching, Form 2 DLP (mean= 4.57) and Form 2 non-DLP's (mean=

4.33) UPSR English results were not significantly different, given that t(38)= 1.848, p= 0.072. Therefore, subjects' prior English proficiency could finally be confidently assumed to be comparable, hence controlled for. At this stage, the data was ready for research question 1.

3.12.2 Preliminary Data Analysis for Research Question 2

Table 3.8

Correlations between Subjects' Vocabulary Performance and English Tuition

		Vocabulary Score	English Tuition
Vocabulary Score	Pearson Correlation	1	.056
	Sig. (2-tailed)		.681
English Tuition	Pearson Correlation	.056	1
	Sig. (2-tailed)	.681	

Note: Correlation is significant at the 0.05 level (2-tailed).

To prepare for research question 2, the researcher needed to investigate subjects' extracurricular English tuition hours to find out if it possessed any threat to the validity of study. For that reason, Pearson correlation test was conducted. The correlation between subjects' English tuition and academic vocabulary performance shown in Table 3.8 affirmed that subjects' English tuition could be ruled out as a confounding factor because of the negligible correlation, $r(57)=.056$, $p=0.681$; hence there was no need to control for this variable and the data was ready for research question 2.

3.13 Subjects' Profile

Table 3.9

Descriptive Data of Finalized Subjects' Demographics

		Gender		Total	Primary School		Total	First language		Total
		Male	Female		SK	Boarding		BM	Others	
School Group	F2 DLP	6	16	22	22	0	22	21	1	22
	F2 Non-DLP	5	13	18	14	4	18	18	0	18
	F4-Non-DLP	9	26	35	33	2	35	38	0	35
	Total	20	55	75	69	6	75	74	1	75

As a summary, this study involved one DLP school and one non-DLP school (namely School A and School B) in the same district. They are 8km or 17minute-drive apart. Both schools were made up of a majority of Malay-speaking students who spoke Malay language at home. All subjects were the best performing students in their respective schools and grades. Throughout the study, the research came into contact with a total of 109 students before any elimination process. After discounting students who spoke English at home and those without consent forms, the total was reduced to 92 subjects. Lastly, after controlling for subjects' UPSR English results, the final number of qualified subjects was 75; their demographics and background are summarized in Table 3.9 below. This group of subjects consisted of 22 Form 2 DLP, 18 Form 2 non-DLP and 35 Form 4 non-DLP students (of which 20 were males and 55 were females). Majority of them (92%) attended Malay-medium national school and almost all (except one Tamil-speaking student) considered Malay language as their first language.

3.14 Conclusion

This chapter began by establishing a positivist quantitative paradigm as the research philosophy adopted for this study which then dictated its succeeding methodology. Three instrument choices: academic vocabulary test, questionnaire and semi-structured interview were then discussed. After that data collection procedure, ethical considerations, and steps to enhance validity and reliability of this study were also presented. Chapter 3 culminated in data analysis, including preliminary data analysis to check for confounding factors (i.e. UPSR English result and English tuition), paving the way for presentation of results to answer the three research questions in the successive chapter.

CHAPTER 4

DATA ANALYSIS

4.1 Introduction

Considering the current state of CLIL research with its methodological flaws (Bruton, 2011 & 2013; Dallinger et al., 2016a; Pladevall-Ballester & Vallbona, 2016), this study was designed with a clear intent to control for these three confounding factors: *selection effect*, *total instruction hours in English* and *extramural English exposure* before any claims were made about CLIL and its impact on students' academic vocabulary learning. The first part of this chapter presents quantitative findings from inferential statistics related to each of the research questions:

1. Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Two non-DLP groups given that the selection effect is controlled?

2. Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Four non-DLP groups given the total instruction hours in English are controlled?

3. Is there any difference between DLP and non-DLP students' Extramural English Exposure? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English a confounding factor in this study?

The second part of the chapter reports on the qualitative data from the interview for the purpose of triangulation to add depth to the later discussion. Finally all results are recapitulated in the conclusion.

4.2 Research Question 1

Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Two non-DLP groups given that the selection effect is controlled?

Table 4.1

Descriptive Statistics and Independent T-test on Academic Vocabulary Performance Form between Form 2 DLP and Form 2 Non-DLP

Group	N	Mean	Standard deviation	Levene's Test for Equality of Variances		t-test for Equality of Means (equal variances assumed)		
				F	Sig.	t	df	Sig (2-tailed)
Form 2 DLP	22	18.27	5.365	0.390	.536	1.230	38	.226
Form 2 Non-DLP	18	16.28	4.763					

* p< 0.05 level of significance

To determine whether Form 2 DLP outperformed Form 2 non-DLP in the academic vocabulary test, an independent t-test was applied to compare the mean scores of both groups. Before independent t-test could be conducted, it was important to first fulfil assumptions required by independent T-test. The normality and homogeneity of variances were checked using Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test (p=0.195, p >0.05) and a visual inspection of box plot showed that the academic vocabulary scores were normally distributed with a skewness of -0.233 (SE = 0.277) and a kurtosis of -0.730 (SE = 0.548). Besides, the Levene's test indicated equal variances (F=0.390, p=0.226), showing that the assumption for homogeneity of variance was fulfilled.

As shown in Table 4.1, Form 2 DLP had a higher mean score in the academic vocabulary test (mean= 18.27) compared to Form 2 non-DLP (mean= 16.28). However, the t-test showed that the difference was not significant, given that t(38)=1.23, p=0.226, at the 5% level of significance. In other words, Form 2 DLP did not outperform Form 2 non-DLP when selection effect was controlled.

4.3   Research Question Two

Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Four Non-DLP groups given the total instruction hours in English are controlled?

Table 4.2

Descriptive Statistics and Independent T-test on Academic Vocabulary Performance Form between Form 2 DLP and Form 4 Non-DLP

				Levene's Test for Equality of Variances		t-test for Equality of Means (equal variances assumed)		
Group	N	Mean	Standard deviation	F	Sig.	t	df	Sig (2-tailed)
Form 2 DLP	22	18.27	5.365	1.050	.310	-1.103	55	.275
Form 4 Non-DLP	35	19.69	4.255					

* p< 0.05 level of significance

Referring to Table 4.2 above, even though by looking at the descriptive data Form 4 non-DLP (mean=19.69) seemed to have a higher average than Form 2 DLP (mean=18.24), the difference was shown to be not significant, t(55)=-1.1, p=0.275. Therefore, when the total instruction hours in English were controlled for both groups, Form 2 DLP subjects did not outperform their Form 4 non-DLP counterparts.

4.4   Research Question 3

Is there any difference between DLP and non-DLP students' Extramural English Exposure? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English exposure a confounding variable in this study?

Table 4.3

Anova Result for Overall Extramural Exposure

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	112.972	2	56.486	1.191	0.310
Within Groups	3415.615	72	47.439		
Total	3528.587	74			

* $p < 0.05$ level of significance

Table 4.4

Correlations between Subjects' Vocabulary Performance and Extramural Exposure

		Vocabulary Score	Extramural Exposure
Vocabulary Score	Pearson Correlation	1	-.109
	Sig. (2-tailed)		.681
Extramural Exposure	Pearson Correlation	-.109	1
	Sig. (2-tailed)	.681	

Note: Correlation is significant at the 0.05 level (2-tailed)

Research question 3 came in three parts in order to determine whether subjects' extramural English exposure was actually a confounding variable that needed to be controlled. Fourteen items in the questionnaire were combined into one variable, namely 'total exposure' to measure 'extramural English exposure' construct. One Way ANOVA was then employed to first examine whether if there was any significant

difference in extramural English exposure within these 3 groups (Form 2 DLP, Form 2 non-DLP and Form 4 non-DLP).

Before One-way ANOVA was conducted, assumptions of normally distributed data and homogeneity of variance needed to be fulfilled. One-sample Kolmogorov-Smirnov Test showed p=0.74, which was more than 0.05, so the 'total exposure' data was said to be normally distributed. As for the Levene's test of homogeneity of variances, the F value of Levene's test was 0.134 with p= 0.875 so equal variances were assumed.

The analysis in Table 4.3 shows that DLP and non-DLP groups did not exhibit any significant difference, such that F (2, 74) =1.19, p= 0.310, at the 5% level. In other words, DLP and non-DLP groups were similar and comparable when it comes to English-related extracurricular activities. Apart from that, it was further confirmed that extramural English exposure was not a threat to the study when the correlation between 'total exposure' and academic vocab score (refer to Table 4.4) was shown to be negligible, r= -.109, p>.05. As a result, the possibility of extramural English being a confounding factor in this study was ruled out.

4.5   Qualitative Data from Interview

Based on research questions 1 to 3, when three common confounding variables (selection effect, total English instruction hours and extramural English exposure) were controlled, DLP students were unable to surpass their non-DLP peers in terms of their academic vocabulary performance. The interview with a DLP teacher, Cikgu Maryam, served as a triangulation to provide a more accurate and valid understanding about DLP through her description of the real-life DLP operation on the ground. The result of the interview was categorized into 5 themes along with transcript excerpts

from Cikgu Maryam. These insights were important in explaining the reasons behind

DLP's apparent lack of effectiveness in the last chapter.

4.5.1 Attitude and Perception of Role as DLP Teacher

Cikgu Maryam seemed to have a clear understanding about the dual nature of DLP

which concerns mastery of both subjects like Math and Science as well as English

language.

> "*DLP is encouraged by the ministry of education encourage the students to learn Math and Science in English. The purpose is to strengthen their communication skills*".

She also believed that language and content learning and teaching go hand in

hand. She took on the role of a language teacher to teach certain language items in

class sometimes. Therefore, contrary to the common stereotype stated in May (2011)'s

study about PPSMI teachers seeing themselves as 'only subject teacher', Cikgu

Maryam did not dichotomize language and content learning.

> *"Yes. [I see myself having responsibility] I am not 100% proficient in English, but I will make sure they speak fully in English…Both actually go hand-in-hand. On the one hand I teach Math, on the other hand I use English words…I don't separate them."*

> *"I also taught them about present and past tense. For example, 'understood', 'understand'…So I teach English occasionally. Most of the time, it will happen naturally but now I will put it in my lesson plans 'English learning' for today. For example, I teach them synonym and antonym. I will also include a bit of vocabulary lesson in my induction set."*

4.5.2 Challenges DLP Teacher faced

When asked about the challenges she faced in DLP class, Cikgu Maryam shared her

struggle when she first started teaching DLP because her students were not accustomed

to having the subjects in English after going through 6 years of primary education in

Malay language. From here, it can be deduced that for DLP to take effect, it might

entail a longer time than just a year because students needed time to warm up and settle

into the DLP environment.

"So during the first or two months, it was really a struggle to teach them…They are used to Malay, to change the terms to English, it is really difficult at first…I used both languages…then gradually reducing Malay by using more English."

Besides, DLP teachers' language proficiency could be an obstacle to the success of DLP. Cikgu Maryam confided to the researcher several times throughout the interview that she did not think she was the language expert in class, but rather she found herself co-learning with her students.

"I think my vocabulary also got improvement".
"If there's any vocabulary or grammatical errors…actually I am not that good at English…Students sometimes said, 'teacher, actually just now you made a mistake'…So it was a two-way communication…because this class most of them got A in English, so they are good at grammar. So we learn together."

4.5.3 Expectation to Teach English and Pedagogical Support

It was also found out that there was no formal expectation for DLP teachers to have language objectives in their lessons. Therefore, it was up to individual teachers how much English exposure was available to their students and each DLP classroom could be very different−some teachers might be more willing to incorporate language lessons than the others.

"There is no formal KPI. They only gave pedagogy and asked us to do it. But there is no proper standard set for us. It is like implicit."

Moreover, while Cikgu Maryam had attended many trainings, pedagogical support for teachers to effectively integrate content and language in their lessons was scarce. In fact, most training courses that she attended were more similar to English language courses to raise teachers' language proficiency than CLIL-related training.

"I had to go for a 2-day training. Everything is related to vocabulary, nothing to do with Math or Science. There's grammar, vocabulary, essay writing. Then we have to communicate in English, do dialog, it's more or less like MUET…There's group presentation. We have to teach to other teachers from other district in English. Teachers have to get the certificate in order to be qualified to teach DLP."

4.5.4 *Classroom Today*

From the interview, it was not difficult to sense that classroom today is shifting from traditional blackboards or "listen and learn" approach to a 21st century skills-oriented classroom where students play a more active role in taking ownership of their learning. Therefore, DLP today is very different from PPSMI a decade ago because the general classroom culture in Malaysia has transformed. From her description, Cikgu Maryam's typical lessons seemed to be filled with student-fronted, engaging activities.

> "For PPSMI, we just used the traditional method, chalk and talk learning only. But now in DLP there are many ways to teach and learn. We can go to the computer lab. I show them slides, they will learn using apps…So many interesting ways."

> "My typical lesson starts when I enter the class, they will be in groups of 5. For me, I don't explain only. I will give each group different questions, so they will discuss amongst them then the group will present on the board. I really want to train them, not just me doing the talking all the time like the traditional method. If they got it wrong, then I will explain the lesson…I used gallery walk so that they will talk to their friends in English to improve their communication skills."

4.5.5 **English Learning Opportunities**

With regards to how she thought her students were learning English in her class, Cikgu Maryam believed language learning would happen naturally in the English-speaking environment especially when they were involved in activities that required them to communicate in English.

> "The students they can communicate in English, they can understand in terms of math in English."

Therefore, even though she felt responsible for her students' English language skill development, there was not a systematic plan for teaching English in her DLP lessons.

> "If I remember, I will teach them. It depends on the topic, if there is anything, I will remind them."

Besides, Cikgu Maryam incorporated technology most of the time into her lesson, not only to aid students' learning of Math, but she believed that use of technology helped with students' language learning indirectly.

> *"When I struggle to make them understand, I use technology. I asked them to bring hand phones to school. Everything in the apps is in English so they will be more familiar with English terms"*

> *"In my class, I teach them to use math applications…so in the apps, the buttons, the instructions are all in English. This is one of the steps to improve their English. Indirectly, they are learning English."*

4.6   Conclusion

This chapter presented the major findings for the research. It began by showing the results of the data analysis according to each research question. Both quantitative and qualitative results were showcased.

To find out ultimately if CLIL is beneficial for academic vocabulary learning, the present study set out to compare DLP and non-DLP students' academic vocabulary performance when confounding factors such as selection effect, total English instruction hours and extramural English exposure were controlled. The results of the inferential statistics indicated that when these confounding factors were stringently controlled, DLP students did not outperform non-DLP students such that there was no significant difference in their academic vocabulary test score. At this point, it might seem that DLP did not contribute to students' academic vocabulary learning.

The qualitative findings from interview were then also presented to explore possible factors that could lead to such result. The last chapter will put forward in-depth discussions of the results, followed by implications and conclusion of the current study.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1   Introduction

The present study aims at finding out if CLIL contributes to academic vocabulary learning when major confounding factors are controlled. The first portion of the chapter attempts to summarize the research finding according to each research question. The middle portion of the chapter is largely dedicated to propose some possible explanations to explain the results of the study, followed by the implications thereof. The last part of the chapter presents the limitations of the study and recommendations for future research.

5.2   Summary of Key Findings

The review of CLIL literature has elucidated methodological flaws that are prevalent across CLIL studies pertaining to language learning. With these common overlooked predictors- prior language proficiency, total formal English instruction time and extramural English exposure controlled, the research questions could be answered in an impartial manner.

5.2.1   Research Question 1

Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Two non-DLP groups given that the selection effect is controlled?

When the selection effect was controlled, Form 2 DLP and Form 2 non-DLP showed no significant difference in their academic vocabulary test scores. DLP's superiority was immediately stripped when both groups' prior English proficiency were matched. It is interesting to point out that although Form 2 DLP actually had extra 280 hours of formal English instruction, they still did not manage to outperform

non-DLP. This validates the existence of selection effect which has skewed the actual CLIL effect in many past studies, hence warranting a more critical re-examination of past CLIL evidences (Bruton 2011, 2013).

5.2.2 Research Question 2

Is there any significant difference in the academic vocabulary performance between Form Two DLP and Form Four non-DLP groups given the total instruction hours in English are controlled?

When total instruction hours in English was controlled, DLP and non-DLP groups again showed no significant difference in their results. In the absence of time advantage, Form 2 DLP did not outperform Form 4 non-DLP. The outcome was unsurprising because in Research Question 1, DLP students even failed to surpass their same-age DLP peers when they clearly had a time advantage, suggesting that having extra instruction time in English during DLP did not benefit students much in their academic vocabulary learning because quality of instruction mattered more than its quantity.

However, it is important to highlight the fact that Form 2 DLP could measure up to the more cognitively matured Form 4 non-DLP while Form 2 non-DLP could not. This shows that DLP group did have some degree of superiority over their same-age non-DLP counterparts though not reflected in the test scores.

5.2.3 Research Question 3

Is there any difference between DLP and non-DLP students' extramural English exposure? Is there any correlation between students' extramural English and their academic vocabulary performance? Is extramural English a confounding factor in this study?

Research question 3 investigated the confounding effect related to extramural English exposure. It was affirmed that extramural exposure was not a confounding factor to CLIL effect because there was no significant difference in both groups' extramural English exposure and no significant correlation between students' extracurricular English-related activities and their academic vocabulary learning. Nonetheless, valuable insights could still be gained from the data and it will be discussed later in the implications section.

5.3   Discussion of Results

The overall result seems to show that DLP students have failed to surpass their non-DLP peers in terms of their academic vocabulary performance. It was congruent with several past studies [e.g. Agustin-Llach (2015); Arribas (2016); Olsson (2015)] such that CLIL groups did not show more outstanding result than traditional ESL groups. It can therefore be argued that CLIL-effect might have been over-appraised to a certain degree in the past studies which were in favour of effectiveness of CLIL in vocabulary learning [e.g. Merikivi & Pietila (2014); Moghadam & Fatemipour (2014); Xantou (2011); Alonso (2015); Lorenzo & Rodriguez (2014); Varkuti (2010); Agustin-Llach (2015); Xanthou (2010)]. That brings into question whether CLIL can actually contribute to academic vocabulary learning as per claimed.

With that said, this is not the writer's intent to negate all the positive claims about CLIL. By judging the research outcome from the surface, it might be tempting to jump into conclusion that CLIL did not "work" in Malaysian classroom and move on to look for a "better" approach that yields more impressive transformation. Nevertheless, annulling CLIL-effect all together just because DLP did not outperform non-DLP is equivalent to throwing the baby out of the bath water. The fact that Form 2 DLP was on par with their 2-year-older Form 4 non-DLP showed potential of DLP which was

unfortunately not fully materialized yet. The lack of recognizable results from DLP might allude to implementation issues rather than the approach per se. In other words, there might be misalignments between theory and de facto CLIL classrooms, resulting benefits and effectiveness of CLIL to be curtailed by various context-specific issues. In this section, four most prominent issues referenced from previous CLIL studies including PPSMI research and the interview data will be presented.

5.3.1 Issue 1: Lack of Structure in DLP-related policy

In most places especially Europe, implementation often lacks structure; as it was already wrestled with in chapter 1, CLIL is a broad term which can mean anything to anyone without clear specifications of what it entails. While Malaysian DLP swings more towards the content-focused spectrum, similar to countries like Hong Kong and Spain, it is still crucial to have clearly-defined perimeters and official regulatory guidelines to explain the 'what's and 'how's of DLP. Without those clearly communicated specifications and expectations, it might give rise of inconsistent DLP implementations across Malaysia. It was the case in Arribas's (2016) study such that irregular CLIL implementation in schools attributed to the lack CLIL impact. However, almost all DLP documents that the researcher encountered did not provide much detail about how DLP should be executed in class but rather about procedures for DLP school application.

Besides, in order for a CLIL program to be truly content and language integrated, it has to have a two-fold objective: learning and teaching of non-language content and language happening at the same time; this was also clearly stated in the DLP circular released by the Ministry of Education (KPM, 2016a). However, by examining closely the latest Dokumen Standard Kurikulum dan Pentaksiran (DSKP) or Standard Curriculum and Instruction which is the national curriculum framework

(KPM, 2016b), it is not difficult to spot the mismatch between DLP's aspiration and national curriculum design. Taking Form 1 DSKP for science as example, it has subject-related goals such as to master scientific knowledge, skills and values in science subject, but language learning and teaching objectives are not mentioned. Cikgu Maryam also corroborated the absence of official key performance indicators (KPI) for subject teachers to teach English and meet any language objectives.

Therefore, when the need to focus on both subject area and language learning is not explicitly stated, it might affect how content-area teachers view the need to take on 'extra' language teacher role due to the lack of accountability, as how Hoare (2011) termed it as one of the policy weaknesses that is common around the world. In his study, immersion teachers were unwilling to consider an immersion agenda because it was not their responsibility, knowing that their performance as a teacher was not judged on this aspect of their teaching. Even though Cikgu Maryam recognized her responsibility to look out for her students' language development, it was her own initiative and other teachers might hold a different view. Hence, the lack of top-down explicit requirement for language matters in DLP class will have immediate implications on how much language learning and teaching is actually taking place.

5.3.2 Issue 2: Teachers Assumed Language Learning Was Always Automatic and Guaranteed.

When Cikgu Maryam was asked if she viewed herself as an English teacher, her response was very different from that of PPSMI teachers in Ong and Tan (2008) and Tan's (2011) studies. Tan's (2011) study reported PPSMI teachers' perceived themselves 'content teacher only' because they were trained as subject teachers, so they did not see the need to shoulder language teaching responsibilities. However, DLP teachers like Cikgu Maryam understood that duality of DLP and did not think language

and content as separate entities. While improvement in could be seen in DLP today in terms of language awareness, the understanding of the role language plays in cognitive development and content learning was still superficial.

Cikgu Maryam disclosed that she used code-switching sometimes to help students understand the topics but the 'terms' remained in English. The researcher does not intend to dwell on the appropriateness of code-switching, but to point out the apparent decontextualized view of language role and truncating it to mere conversion of Malay terms to English terms. Her emphasis on Mathematical terms in English could also be noticed when she shared about how students were learning English Mathematical terms, therefore learning English indirectly through the use Mathematics apps in class. This simplified view of the role language plays in meaning making of academic matters and knowledge construction is common among CLIL teachers (Dalton-Puffer, 2011; Nikula, 2015).

Besides, there was actually occasional inclusion of brief English lessons in Cikgu Maryam's class. This shows how different DLP today is from PPSMI last time because a Math teacher like Cikgu Maryam was taking on the instructive role in English teaching, though on a small scale. However, these efforts are not consistent; there is no vision or larger overall plan to systematically integrate content and language teaching. The topics were random and irrelevant to the content or it was on an ad hoc basis depending what came to the teacher's mind as she was teaching. This might come from the teacher's belief that teaching subject matter itself in the target language will automatically result in learning in both domains, so students can always learn something about English as long as they are in an 'English environment'. However, as how Swain (1988) aptly put, "good content teaching is not necessarily good language teaching...the integration of language, subject area knowledge, and thinking skills

requires systematic monitoring and planning" (p.68). Such random, spontaneous and 'go with the flow' belief is not uncommon in many CLIL studies. Many CLIL teachers believed that language acquisition in immersion was 'primarily incidental' and language development would happen naturally through extended exposure (Walker & Tedick, 2000; Hoare, 2011). Interestingly, Dalton-Puffer and Smit's (2013) 48 interviewed stakeholders also reckoned that students were incidentally 'picking up the language even during unstructured and unorganised CLIL lessons with sporadic teaching and learning in English.

Nonetheless, the lack of calculated plan and deliberate attention to language development during subject matter instruction does not lead to optimum target language acquisition (Cammarata & Tedick, 2012). To be effective, language teaching must be carefully considered and planned. The language teaching can be done with different approaches, either explicitly or implicitly. Therefore it is not the approach that the researcher is discussing here, but the lack of intentional planning to achieve language objectives and unrealistic assumption that language learning is guaranteed in every lesson. If the curriculum and instruction is not developmental from a language-learning point of view, students' language skills will remain stagnant because there is always the risk of repetition and redundancy in everyday instruction without teachers' purposive lesson design. Besides, in the absence of such plans, teachers may provide language learners with inconsistent and random information about target language forms (Genesee, 1994), which do not contribute much to language learning. Therefore, even though Form 2 DLP subjects did have more instruction hours in English than their same-age peers since they had the extra hours of Math and Science in English, the opportunities for their language development were unfortunately taken for granted.

5.3.3 Issue 3: Lack of Expertise in Content-language Integration

The effectiveness of DLP lessons in helping students learn English also hinges heavily on individual teachers' competencies not only in content area, but in English language and most importantly in how language and content should be integrated. Unfortunately immersion teachers were often described as ill-prepared to bring about language learning during content lessons or even to recognize students' language-learning needs (Met & Lorenz, 1997), not because they were unwilling to take on language teacher role but didn't know how curricular content, concepts and facts should be meaningfully integrated with lexical and grammatical aspects; neither did they know what tasks were suitable (Cammarata & Tedick, 2012; Tan, 2011). Finding the exact language embedded in content to guide students' attention and aligning content objectives with language objectives require different skill sets, experience and expertise. This is when teacher training and continuing professional development come into play.

During PPSMI time, "Buddy System" was established with the goal of having English teachers as language resource persons to support the subject teachers in the same school. However, this program was encumbered by the fact that the English teachers were unable to help Math and Science teaches much because of their limited content-area knowledge (Tan 2011; Paran 2013). Learning the lesson well from PPSMI, many years after, according to Cikgu Maryam, there have been many training courses provided for DLP teachers.

Not only DLP teachers, all teachers today are required to attend training courses rather frequently. Especially in these recent few years, the 21st century learning skills has been steadfastly championed by the Ministry of Education hence gaining popularity among educators in Malaysia. Compared to teachers a decade ago, teachers

today receive much more training on general pedagogical principles and hands-on techniques in order to equip students with certain core competencies such as collaboration, digital literacy, critical thinking, and problem-solving to prepare students for today's world. Cikgu Maryam found pedagogy-focused training like this complementing DLP very well; she had learned many techniques that she could apply in her class on a daily basis such as Round Robin, Think-Pair-Share, etc.

When asked about DLP-specific training, according to her, all courses she had attended then were more geared towards English language training to boost DLP teachers' English proficiency. Even though general pedagogical training and language courses are beneficial to DLP teachers, when it comes to delivering language objectives embedded in rich content, it can still be challenging, especially when CALP is concerned. The idea of academic English is usually foreign to content teachers and is often limited to content-specific words. When there is no CLIL-specific pre-service preparatory courses and in-service professional development focused on language and content integration, it handicaps teachers' ability to give quality CLIL instruction hence missing opportunities for students' language acquisition (Paran, 2013).

From the interview, it can also be inferred that teachers' English proficiency continues to be a challenge even years after PPSMI because language courses are still needed for DLP teachers. During PPSMI time, subject teachers were trained to teach Math and Science in Malay in teachers colleges. Many PPSMI teachers admitted difficulties especially in their efforts to convey concepts and ideas to their students verbally (Ong & Tan, 2008) such that Juriah Long and colleagues' ("Poor English impedes lessons", 2006) study revealed that 55% teachers felt uneasy using English in class. A young teacher like Cikgu Maryam experienced PPSMI not as a teacher, but as a student. In university, she was trained in English for every Mathematics subject.

However, the lack of proficiency persists even among younger teachers like her. Throughout the conversation, she mentioned several times that she was not proficient in English, especially with regards to grammar, and she did not think herself as the language expert in the class. Therefore instead of modelling the language use to show her students different language patterns and to encourage students to imitate, she regarded it as a mutual learning process, *"sama-sama"* according to her. When asked if she gave any form of feedbacks on students' language use in class, she claimed that she would do it sometimes, but because of her lack of language competency, there were times her students would correct her instead. Even though teacher-student mutual learning process is commendable, if teachers are not proficient in the medium of instruction, when it comes to exploring scientific concepts in a more cognitively challenging manner, being the facilitator in inquiry-based investigations, or engaging students in high level questioning, many opportunities to grow students' language repertoire might not be taken full advantage of (Mohd Satt & Othman, 2010).

5.3.4 Issue 4: Lack of CALP Development Opportunities

Having experienced PPSMI herself as a student, Cikgu Maryam could spot several differences between PPSMI and DLP especially in terms of pedagogies and the incorporation of technology in the class. Different from stereotyped traditional classrooms, her DLP classroom was usually packed with activities that she learned from the 21st century learning skills trainings she attended. Although it is a good thing that the traditional teacher-centered chalk-and-talk style of teaching is gradually taking a back seat, abundancy of exposure to language quantified by number of in-class activities does not always warrant quality CLIL instruction. This is because subject teaching in an L2 does not in itself guarantee communicativeness of L2, as ultimately it is the pedagogy used by the teacher that will generate communication (Paran, 2013).

This explains why extra instruction hours did not grant Form 2 DLP a clear advantage in lexical competence over their Form 2 non-DLP peers since it is not just about extending exposure to L2, but quality thereof matters even more. This section highly two blind spots that are commonplace germane to CLIL classroom instructions.

Firstly, due to the naturalistic theoretical assumptions about language learning, CLIL is mostly associated with implicit learning that provides students with exposure to the target language in an authentic and meaningful context. Likewise, DLP's general orientation to language remains implicit rather than explicit; it is believed that language learning through incidentally without too much concern for correctness or emphasis on form because it is the message they're conveying that matters more. A classroom like this highly depends on teachers' provision of linguistic input through different means to ensure sufficient and quality exposure to the target language for incidental language acquisition to take place. Nonetheless, Dalton-Puffer's (2007) [cited in her 2011's work] analyses of 40-class discourses across 14 English CLIL classrooms in Austria revealed that students were often not provided with sufficient rich linguistic input because teacher utterances were far too short and lacked syntactic complexity. This was the case when Cikgu Maryam admitted her weakness in her language proficiency. Therefore, even though a classroom can be activity-filled, the lack of language-sensitive input (which is supposed to be a distinctive feature of a CLIL classroom to expand their linguistic repertoire) is one of the potential stumbling blocks for DLP to attain its full potential.

Secondly, even though Cikgu Maryam insisted on having English only in class, CALP acquisition is not guaranteed because communication could easily be done with BICS without sustained, extended discourse or being cognitively engaged. It was the case in Hajer's (2000) study that the teachers tended to use everyday vocabulary in

explaining the subject matter, instead of seizing the opportunity to apply formal academic vocabulary. When the output is content-oriented and activities do not explicitly promote students' linguistic development and application of academic language, the opportunities to stimulate conceptual and CALP development are not optimized. Besides, in Cikgu Maryam's DLP class, the focus of language seemed to be limited to subject-related terms. She also perceived that Mathematics was mostly about calculations and using the correct terms, unlike Science which usually came in long English sentences. Furthermore, a study conducted by Yassin, Ong, Alimon, Baharom, and Lai (2010) about learners' cognitive engagement and teacher questions revealed that teachers frequently used low-level questions even though students in the observed classes were sufficiently proficient in English to process language and content more deeply. The same problem might persist in DLP class especially when DLP teachers' lack of language proficiency limits their ability to facilitate more cognitively demanding interactions in more complex language. This brings to question the "how much is cognition-language integration really taking place in CLIL classroom?". The predominance of lower level thinking tasks in DLP classrooms might be a threat in hampering students' CALP learning. Even with the incorporation of educational technological resources in Cikgu Maryam's lessons, although it is commendable, CALP will only take effect in such CLIL environment if technology-based learning is thoughtfully geared towards application of higher order thinking skills and academic language which is in nature more complex, hence demanding more sophisticated tasks. Otherwise, there is always a risk that the lesson may appear to be interactive but in reality it is still transmissive in style, with a strong focus on mere facts and technical terminology.

The above highlights the paucity of CALP-sensitive pedagogy as a barrier to learners' cognitive as well as academic language development. Although the DLP classroom is lively and engaging, students might not benefit much from it from language perspective because of insufficient meaningful inputs and CALP application opportunities.

5.4   Implications

The issues above resonates with Wolff's (2009) (cited in Meyer et al, 2015) claim on the unrealized full potential of CLIL. In other words, a deeper integration of content and language has not yet been fully materialized in DLP classrooms. This section discusses eight pivotal measures informed by past literature that could potential help improve the effectiveness of DLP implementation. The researcher definitely does not claim to have all the know-how needed to make a CLIL program successful and by no means offer a simple solution to CLIL implementation, but it does extend an invitation to explore and examine areas for improvements.

5.4.1   Implementation Guideline Needed

First and foremost, a more thorough and accessible implementation policy for DLP is indispensable to clearly communicate to stakeholders of different stages its institutional goals to ensure uniform and consistent implementation of DLP across the board. Sylvén (2013) affirms that an all-inclusive guideline is key to effective teaching so that CLIL teachers can be trained accordingly to integrate language and subject learning in standardized manner. Without which, it is feared that DLP will eventually become just another buzzword, if implementers are uncertain about the must-have appropriated characteristics of DLP to differentiate it from generic mainstream classrooms. It is also highly possible that not every DLP teacher grasps the duality of DLP as well as Cikgu Maryam. Therefore, there is a dire need for comprehensive

policy and guidelines that appeal to local education ecosystem. Once the big picture is set in place, it forms a backbone for stakeholders of different levels to work together on bringing out the best of DLP.

Apart from that, even though Malaysian CLIL program is content-focused, having explicit mention of language focus and goals is necessary. It is to hold language outcome the same level of accountability as content achievement so that teachers are motivated to engage in the more difficult undertaking of teaching both content and language so that the dualistic nature of DLP goal can be consistently actualized in classroom. Then this content-language integration can finally be effectively systematized and reflected in every tier in the system, from DKSP and DLP textbooks to teacher education, pedagogy and classroom practices.

5.4.2 DLP-specific Training Needed

Once the official call for attention to DLP students' language development is set in order, it will have implications on DLP teachers' training as much as classroom practice. Mohd Satt and Othman (2010) called for an overhaul and restructuring of content teacher education programs in their PPSMI study, but the appeal remains the same today. There should be CLIL-specific pre-service and on-going professional development to continually equip DLP teachers with skills to develop language-specific objectives for their daily classroom practice, but most importantly for DLP teachers to be more capable of integrating language and content instruction functionally and meaningfully in order to promote CALP development.

5.4.3 Better Understanding of Language and Content Interdependence

However, all these will not suffice if DLP teachers' over-simplistic perception towards language learning, assuming that language learning will take care of itself as long as students are in an English-speaking environment, remains unchanged. Even though

DLP teachers today are more willing to taking on language teacher role than PPSMI teacher, the overall orientation regarding language learning is still rather passive. On top of that, the role of academic language in subject learning is poorly understood. In order for DLP to fulfill its CLIL potential, there must be a true integration of language and content, in order words, strong interdependence between content and language. The conceptual separation of CALP from BISC by Cummins (1999; 2008) is useful for teachers to understand having peer appropriate conversational fluency will not suffice for students without grade-appropriate academic language to succeed in school setting. A DLP teacher therefore needs to realize that deep academic understanding cannot happen without reasonable CALP; language is getting complex at higher level because as the cognitive demands of academic content grow, so do the linguistic demands. Therefore if students are left without provision of proactive language assistance, they will not be have the access to content mastery. Therefore, though the primary objective of DLP can be content mastery, DLP teachers need to acknowledge pivotal role language plays in facilitating and enabling cognitive understanding, and academic language is much more than only technical vocabulary (Scarcella, 2003).

5.4.4 Necessity of Thoughtful Planning

With that said, it alludes to the utmost importance of having systematic planning in order to find practical ways to weave cognitively appropriated language objectives into their instructions (Snow, Met, & Genesee, 1989; Tan, 2011; Walker & Tedick, 2000). Having plans helps teachers select appropriate academic tasks and instructional strategies so that both teachers and students are compelled to use these designated language skills and aspects. This is to curb the common 'go-with-the-flow' attitude, expecting students to 'pick up' target language in a desultory fashion.

5.4.5 More Quality Language Input by Teachers

CLIL is often known for its natural approach (Krashen & Terrell, 1983) in which language learning is approached implicitly via creating authentic context, non-stressful environment for learners to pick up language incidentally. This is made possible through rich comprehensible input, based on Krashen's input hypothesis (1982), which is one step beyond a learner's current linguistic abilities (i+1). However, due to teachers' limited proficiency in English and the preceding unhelpful assumption about 'osmosis-style' language learning, a classroom might full of activities, however, it might not be necessarily be language-rich. DLP classrooms have potential to provide abundant comprehensible input especially when students are introduced to new and complex linguistic systems, content provides cognitive basis for language learning. Therefore, to imbue students with meaningful and authentic inputs, it is beyond merely conducting activities in English. Teachers need to be competent and conscious of the language inputs they provide students with, so that they are comprehensible at the same time cognitively-stimulating, not wanting to rely on the proficiency students already have. For CLIL to truly achieve its meaning-based approach objectives, what teachers have to do must be more than just using English in class at all time.

De graff et. al.'s (2007) CLIL teaching observation tool which consists of several indicators for meaningful exposure is a comprehensive guideline for DLP teachers to select and tailor input material in order to make it challenging but comprehensible enough for learners. This observation tool ensures meaningfulness beginning from meaning-focused input (such as text selection and teacher talk), then to meaning-focused processing (which focuses on stimulating meaning identification), and lastly to meaningful use of English (e.g. eliciting responses). Therefore, for DLP to be effectively meaning-focused, mere exposure to language is not enough and if the

input is not processed for meaning. Furthermore, learners cannot be left to grapple meaning alone because it can be erroneous, so meaning-focused process needs to be facilitated (Dalton-Puffer, 2011).

5.4.6 Addition of Explicit Instruction or Focus on Form

Even though implicit or meaning-based instruction (no explicit teaching of language items) seems to be the prominent attribute of CLIL, relying on it alone is insufficient to guarantee complete language learning because of its shortcomings. For instance, in wanting to create a stress-free environment, teachers' implicit recasts are ineffective because learners often do not notice the teacher's actual intent to draw the learners' attention to language form because of the primary focus on meaning (Lyster, 2004). Besides, focus on meaning alone might lead to unsubstantiated outputs: short and simple responses are acceptable as long as the message is understood. When it comes to vocabulary learning, Schmitt (2008)'s study showed implicit instruction that required students to focus only on the overall message did not generate enough engagement with individual words for them to be remembered. He went on to review Horst, Cobb, and Meara's (1998) study on incidental exposure through reading that found a discouragingly low pick-up rate of new words, with about only one word being correctly identified out of every 12 words tested. This shows that implicit-only or meaning-based-only instruction has limited effect on vocabulary gain.

Dalton-Puffer, Nikula, and Smit (2010) claimed that among European CLIL there seems to be a gradual but significant movement away from sole reliance on implicit or incidental learning approach to explicit "focus on form" of language. Contrary to common understanding, focus on form is not about stereotypical decontextualized, isolated and mechanistic grammar lessons and drills, rather paying attention to formal, lexicogrammatical aspects of language as carriers of meaning,

emphasizing the relationship of form with its communicative or academic functions (Cammarata & Tedick, 2012). Therefore, instead of opting to focus on either 'form' or 'meaning', it is important to tear down the meaning-form dichotomous barrier and for this reason, Lyster's (2007) "counterbalanced approach", as cited in Cammarata and Tedick. (2012), is particularly helpful and is gaining popular in immersion-related research. Counterbalanced approach advocates giving equal weight to meaning focus and form focus in immersion education, in other words, underscoring the need to bring students' attention to linguistic form in the context of content (meaning-based) teaching.

This approach is consistent with pedagogical framework for immersion advocated by Gibbons (2002) which could potential be adapted by DLP after careful study. Gibbons proposed an analogy using an hourglass to exemplify how the focus of teaching and learning should reorient flexibly throughout a lesson. The wide top of the hourglass represents meaning and knowledge construction (meaning-focused); as the hourglass narrows, the focus shifts to language itself, where students' attention is directed to various lexico-grammatical features. This is a crucial stage where a CLIL teacher employs activities aimed at awareness-raising of specific language forms, at the same time giving explicit corrective feedback (De graff et. al.'s, 2007) so that students grapple with the language after overall meanings being made and the curriculum knowledge being constructed (Gibbons, 2002). Lastly, as the hourglass widens again at the bottom, it is when students apply their knowledge of language aspects by returning to meaning and knowledge construction embedded in various tasks. However, the sequence suggested by Gibbons should by no means be treated as mono-directional. The shift from 'focus on meaning' to 'focus on form' or vice versa should be allowed anytime throughout the lesson. Thus, both form- and meaning-

focused instructions should be used in tandem; they should not be treated as if they are mutually exclusive. As how Schmitt (2008) put it when commenting on vocabulary learning, the most effective way of improving incidental learning is by reinforcing it with intentional learning tasks.

5.4.7 CALP-oriented Activities Needed

CALP is key to academic success. As the academic content becomes more complex, so does its language. Learners cannot fully grasp academic content without CALP; to grow CALP, it involves cognitive skills embedded in academic content. However, it would be rather unsurprising that amongst DLP teachers little distinction is made between CALP and BICS. DLP teachers need to realize when subject lessons are conducted in English, students are already exposed to sufficient BICS; it has not occurred to some if not many that when the tasks and instructions are cognitively undemanding, there is no reason for students to use CALP.

As CALP and academic content mastery are intertwined, to stimulate learners' academic language development, they need to be engaged in tasks that are conducive to their cognitive engagement (Tedick & Cammarata, 2012). The responsibility to ensure meaningful CALP input and use in class falls on DLP teachers to seize the opportunities to model more complex language and conduct instructional activities that demand progressively more complex language skills from their students as they are growing in their academic content knowledge (Genesee, 1994). Successful examples of such can be seen in Dalton-Puffer's (2007) and Lorenzo and Rodríguez's (2014) studies of CLIL classrooms in which many higher order questions and extended discourses were observed and CLIL teachers took advantage of this meaning-rich environment made possible by content integration to require learners to participate in expository discourse, make inferences from given facts, and use academic vocabulary

when expressing abstract ideas. Likewise, DLP teachers should demand students to engage in higher order thinking skills (such as defining, classify, illustrate, contrast, compare, give reasons, summarize, predict, sequence, deducing) through sophisticated content-related activities. By doing so, students are brought into situation that requires deep processing of the language and are in-effect required to wrestle with more complex or elaborate language skills (Lyster & Ballinger, 2011; Clegg, 2007). This is also consistent with Swain's (1985) Output Hypothesis which purports that the process of producing a linguistic output can benefit learners' language learning when learners notice their gap in knowledge of the L2. When DLP students are engaged in higher-order-thinking language activities, they are bestowed opportunities to produce linguistic output that facilitate language processing. When CALP learning as such is made available, DLP can then be regarded as a true CLIL environment.

5.4.8 Extramural English as Supplementary Method

Even though in this study it's found out that extramural English exposure is not a confounding factor because there is no note-worthy relationship between learners' overall extramural English exposure and their academic vocabulary result, valuable insights can be inferred from this set of data. In Sweden, Sylven's (2004) study showed that out-of-school reading behaviour was the strongest predictor amongst all extra-curricular activities; while Malaysian students in this study did read, it did not benefit their academic vocabulary acquisition. Therefore, it can be assumed that having extramural exposure doesn't seem to enhance automatically development of academic vocabulary.

Having mere English-related activities afterschool is insufficient to warrant academic vocabulary learning because the encounter with academic vocabulary is highly dependent on the types and genres of their exposure. From subjects' responses

to short-ended questions in the questionnaire, it seemed that a large percentage of the exposure they had gravitated towards popular media, such as watching Disney movies and Ellen Degeneres's shows, and reading Korean Pop Culture news and blockbuster novels. Only a handful cited educational and academic resources, such as watching National Geographic documentaries, reading Wikipedia for leisure and researching about latest technological gadgets. Therefore, some sources of extramural English benefit students' general vocabulary while others might be rich in academic vocabulary. For instance, students who read comics may pick up new words, but this type of reading material might not contribute to their academic vocabulary learning as much compared to those who like scientific magazines or documentaries.

Hence, it is important to differentiate various sources of extramural exposure so that given the right materials, extramural English exposure is a powerful complement to academic vocabulary learning that is taking placing simultaneously in DLP classroom. One way to do it, suggested by Cikgu Maryam, is to take full advantage of an under-utilized online learning platform available for all government schools called the Frog VLE (Virtual Learning Environment). Cikgu Maryam believed that learning should not be confined within the four classroom walls. Frog VLE is one of the readily available avenues where students' afterschool academic vocabulary learning opportunities can be monitored by teachers via sharing informative yet entertaining videos or reading materials (such as Tedtalks and Encyclopedia Britannica) that help reinforce lexical items that are relevant to what they learn in class. These activities should not be deemed as homework and should not be graded.

5.5   Limitations and Future Research

There are several limitations pertaining to the methodology of this study. Firstly, admittedly there are many other factors that may influence learners' scores on their

academic vocabulary test but the current research raises only three specific issues for investigation. There are other factors such as students and teachers' aptitude and motivation, classroom composition and teachers' characteristics could potentially make a difference to the outcome but they are not controlled because it is close to impossible to control all the unidentified interfering variables (Mackey & Gass, 2005) in a study conducted in a natural classroom in lieu of a fully controlled experimental setting. Even though a thorough 'isolation' of CLIL effect is unattainable in this study, this study provides avenues for further research that look into other variables to piece together an even more 'distilled' CLIL effect.

Besides, the relatively small sample size employed in the study might impose limitations on generalizability. Both location and time triangulations are required in the future to enhance the validity of this study. More interviews should also be conducted with other DLP teachers of different subjects and schools. Apart from that, particular characteristics of the participants (i.e. middle-class Malay-speaking group in Selangor) might not be suitable to represent the entire DLP population in Malaysia. Therefore, any future research that involves a bigger sample size with subjects of more diversified background will boost the generalizability of the study.

Moreover, getting CLIL teachers to comment on their own teaching experience might not be reliable and sufficient because one might not feel comfortable to disclose too much information that is not in favor of personal or school reputation. For this reason, in order to secure the insider knowledge of the site to arrive at a more nuanced understanding of the CLIL, it is a desideratum for external researchers to make actual DLP classroom observations to get a hold of what actually happens inside the classroom. Moreover, a one-shot test like this can be made more meaningful by having

pre- and post-tests to document a more accurate reflection of how much they truly change.

As of the instruments, using both AWL and AVL as standards of reference for academic vocabulary might be worthwhile to strengthen the validity of the results. Lastly, like how Abd Manan, Ali, and Shamsudin (2013) quantified the number of academic vocabulary that made it to the previous secondary school syllabi, same method can be applied to the latest DLP textbooks as a way to gauge how likely students are to pick up academic vocabulary through exposure to DLP material.

This research has its limitations so it does not claim to provide very definitive findings about CLIL and academic vocabulary learning. With that said, this study is still valuable to kick-start on-going research and discussions about CLIL and its effect in second language learning in local context.

5.6   Conclusion

This study attempts to compare the academic vocabulary performance between DLP and non-DLP students in a methodologically rigorous way. The data analysis and discussion displayed that there was no significant difference in the academic vocabulary performance between DLP and their non-DLP peers when confounding factors like selection effect, total instruction hours in English and extramural exposure to English language were controlled. While it is tempting to condemn DLP for its lack of recognizable results, it is difficult to disregard the fact that DLP group did manage to stay on par with senior non-DLP students who were two years older. For this reason, it is believed that there might be misalignments between theory and actual DLP classrooms, handicapping DLP students from reaping its full benefits. Therefore, several implementation issues that might have impeded DLP to achieve its fullest potential were explored. In conclusion, the researcher is

optimistic of DLP's endeavor in Malaysia because if implemented properly, DLP has a lot more to offer when it comes to students' English academic vocabulary learning. Therefore, the researcher suggested some measures that are worth-considering to ensure that CLIL's framework and implementation are well-aligned in DLP classrooms to provide teaching and learning environment that is conducive for English academic vocabulary acquisition.

REFERENCES

Abd. Manan, A., Ali, N. L., & Shamsudin, S. (2013). Does the Malaysian English language syllabus cater to the academic vocabulary needs of secondary school students entering universities?. *Jurnal Teknologi (Sciences and Engineering)*, 65(2), 7-14. DOI: 10.11113/jt.v65.2345

Agustín-Llach, M. P. (2015). The effects of the CLIL approach in young foreign language learners' lexical profiles. *International Journal of Bilingual Education and Bilingualism*, 1-17. http://dx.doi.org/10.1080/13670050. 2015.1103208

Agustín-Llach, M. P., and Alonso, A. C. (2016) Vocabulary growth in young CLIL and traditional EFL learners: evidence from research and implications for education. *International Journal of Applied Linguistics*, 26: 211–227. doi: 10.1111/ijal.12090.

Alonso, A. C. (2015). Receptive vocabulary of CLIL and Non-CLIL primary and secondary School Learners. *Complutense Journal of English Studies*, 23, 59-77. http://dx.doi.org/10.5209/rev_CJES.2015.v23.51301

Alonso, E., Grisaleña, J., Campo, A. (2008). Plurilingual Education in Secondary Schools: Analysis of Results. *International CLIL Research Journal,* 1 (1), 36-49.

Anderson, C. E. (2011). CLIL for CALP in the multilingual, pluricultural, globalized knowledge society: Experiences and backgrounds to L2 English usage among Latin American L1 Spanishusers. *Latin American Journal of Content & Language Integrated Learning*, 4(2), 51-66. doi:10.5294/laclil.2011.4.2.5

Arribas, M. (2016). Analysing a whole CLIL school: Students' attitudes, motivation, and receptive vocabulary outcomes. *Latin American Journal of Content and Language Integrated Learning*, 9(2), 267-292. doi:10.5294/laclil.2016.9.2.2

Arukesamy, K. (2015, May 17). Students losing out due to lack of English skill. *The Sun Daily*. Retrieved from http://www.thesundaily.my/news/1421066

Bailey, A. L., Butler, F. A. & Sato, E. (2007). Standards-to-standards linkage under Title III: exploring common language demands in ELD and science standards. *Applied Measurement in Education*, 20(1), 53-78. http://dx.doi.org/10.108 0/08957340709336730

Bailey, A. L. & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school?. *Language Testing*, 28(3), 343-365. DOI: 10.1177/0265532211404187

Banegas, D. L. (2011). Content and Language Integrated Learning in Argentina 2008 – 2011. *Latin American Journal of Content & Language Integrated Learning*, 4(2), 33-50. doi:10.5294/laclil.2011.4.2.4 ISSN 2011-6721

Baumann, J. F. & Graves, M. F. (2010). Commentary: What is academic vocabulary? *Journal of Adolescent & Adult Literacy*, 54(1), 4-12. doi:10.1598/JAAL.54.1.1

Bostwick, M. (1998). The application of immersion education in Japan: Results, challenges and possibilities. In M. R. Childs & R. M. Bostwick (Eds.), *Learning through two languages: Research and practice*. (pp. 8-22). Numazu: Katoh Gakue.

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, 39,523-532. doi:10.1016/j.system.2011.08.002

Bruton, A. (2013). CLIL: Some of the reasons why … and why not. *System*, 41, 587–597. https://doi.org/10.1016/j.system.2013.07.001

Būdvytytė-Gudienė, A. & Toleikienė, R. (2008). Content and Language Integrated Learning: Features of Educational Methods. *Social Sciences*, 60, (2), 90-97.

Cammarata, L., & D. K. Tedick. (2012). Balancing content and language in instruction: The experience of immersion teachers. *The Modern Language Journal, 96(2),* 251-269.

Carlo, M. S., August, D., & McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, 39(2), 188-215.

Catalán, R. M. J., & De Zarobe, Y. R. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL instruction. *Content and language integrated learning: Evidence from research in Europe*, 81-92.

Cenoz, J. (2015). Content-based instruction and content and language integrated learning: the same or different?. *Language, Culture and Curriculum*, 28(1), 8-24.

Cenoz, J. Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. Applied Linguistics, 35, 243–262. doi:10.1093/applin/amt011

Cenoz, J., & Ruiz de Zarobe, Y. (2015). Learning through a second or additional language: content-based instruction and CLIL in the twenty-first century. *Language, Culture and Curriculum*, 28(1), 1-7. DOI: 10.1080/07908318.2014.1000921

Clegg, J. (2007). Analysing the language demands of lessons taught in a second language. *Revista española de lingüística aplicada*, 1, 113-12.

Coyle, D. (2007). Content and language integrated learning: Towards a connected research agenda for CLIL pedagogies. *International journal of bilingual education and bilingualism*, *10*(5), 543-562. http://dx.doi.org/10.2167/beb459.0

Coyle, D. (2008). CLIL—A pedagogical approach from the European perspective. In *Encyclopedia of language and education* (pp. 1200-1214). Springer US.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Coxhead, A. (2011). The Academic Word List ten years on: Research and teaching implications. *TESOL Quarterly*, 45(2), 355-362.

Creswell, J. W. (1994). *Research design: Qualitative & quantitative approaches*. Thousand Oaks, Calif: Sage Publications.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL quarterly*, 175-187.

Cummins, J. (1981). Bilingualism and Minority-Language Children. *Language and Literacy Series*. The Ontario Institute for Studies in Education, 252 Bloor Street West, Toronto, Ontario M5S 1V6.

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

Cummins, J. (1999). *BICS and CALP: Clarifying the Distinction.* Retrieved from https://eric.ed.gov/?id=ED438551

Cummins, J. (2008). BICS and CALP: Empirical and Theoretical Status of the Distinction. In Street, B. & Hornberger, N. H. (Eds.). *Encyclopedia of Language and Education, 2nd Edition*, Volume 2: Literacy. (pp. 71-83). New York: Springer Science + Business Media LLC.

Dallinger, S., Jonkmann, K., Hollm, J., & Fiege, C. (2016a). The effect of content and language integrated learning on students' English and history competences–Killing two birds with one stone?. *Learning and Instruction*, 41, 23-31.

Dallinger, S., Jonkmann, K., & Hollm, J. (2016b). Selectivity of content and language integrated learning programmes in German secondary schools. *International Journal of Bilingual Education and Bilingualism*, 1-12. http://dx.doi.org/10.1080/13670050.2015.1130015

Dalton-Puffer, C. (2008). Outcomes and Processes in Content and Language Integrated (CLIL) Learning: Current Research in Europe. In: Delanoy, W., Volkmann, L. (Eds.), *Future Perspectives in English Language Teaching (pp. 7-23)*. Heidelberg: Carl Winter.

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles?. *Annual Review of applied linguistics*, 31, 182-204.

Dalton-Puffer, C., & Smit, U. (2013). Content and language integrated learning: A research agenda. *Language Teaching*, *46*(4), 545-559.

Dalton-Puffer, C., Nikula, T., & Smit, U. (2010). Charting policies, premises and research on Content and Language Integrated Learning. In C. Dalton-Puffer, T. Nikula and U. Smit (eds.), *Language Use and Language Learning in CLIL Classrooms* (1-19). John Benjamins, Amsterdam.

De Diezmas, E. N. M. (2016). The impact of CLIL on the acquisition of L2 competences and skills in primary education. *International Journal of English Studies*, *16*(2), 81-101.

de Graaff, R., Koopman, G. J., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in content and language integrated learning

(CLIL). *International Journal of Bilingual Education and Bilingualism*, 10(5), 603–624.

Deák, G.O. (2014). Interrelations of language and cognitive development. *Encyclopedia of Language Development* (pp. 284-291). P. Brooks & V. Kampe, Eds. SAGE

Eurydice. (2006). Content and language integrated learning (CLIL) at school in Europe. Retrieved from http://www.indire.it/lucabas/lkmw_file/eurydice/CLIL_EN.pdf

Francis, D. J., Rivera, M., Lesaux, N. K., Kieffer, M. J., & Rivera, H. (2006). Practical guidelines for the education of English language learners. In *Presentation at LEP Partnership Meeting, Washington, DC.* Retrieved from http://www. centeroninstruction. org.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35(3),* 305-327.

Genesee, F. (1994). Integrating language and content: Lessons from immersion. *Educational Practice Reports.* Retrieved from https://escholarship.org/uc/ item/ 61c8k7kh

Genesee, F., & Lindholm-Leary, K. (2013). Two case studies of content-based language education. *Journal of Immersion and Content-Based Language Education*, *1*(1), 3-33.

Gibbons, P. (2002). Learning language, learning through language, and learning about language: Developing an integrated curriculum. In *Scaffolding Language, Scaffolding Learning: Teaching Second Language, Learners in the Mainstream Classroom*. pp 118-138. Portsmouth, NH: Heinemann.

Hajer, M. (2000). Creating a language-promoting classroom: Content-area teachers at work. In J. K. Hall & L. S. Verplaetse (Eds), *Second and foreign language learning through classroom interaction (*265-286). London: Routledge.

Harji, M. B., Balakrishnan, K., Bhar, S. K., & Letchumanan, K. (2015). Vocabulary Levels and Size of Malaysian Undergraduates. *English Language Teaching*, *8*(9), 119.

Hoare, P. (2011). Context and constraints: Immersion in Hong Kong and Mainland China. In D. J. Tedick, D. Christian & T. W. Fortune, *Immersion education: Practices, policies, possibilities* (211-230). Bristol: Short Run Press Ltd.

Jala, I. (2015, November 9). Malaysian survey goes viral, 90% want standard of English raised. *The Star Online.* Retrieved from http://www.thestar.com.my /business/business-news/2015/11/09/aligning- the-constellations/

Juan-Garau M., Prieto-Arranz J.I., Salazar-Noguera J. (2015) Lexico-Grammatical Development in Secondary Education CLIL Learners. In M. Juan-Garau & J. Salazar-Noguera (eds). Content-based Language Learning in Multilingual Educational Environments. *Educational Linguistics* (1790195). Cham: Springer.

Kaur, S., & Sidhu, G. K. (2014). Evaluating the critical literacy practices of tertiary students. *Procedia-Social and Behavioral Sciences*, *123*, 44-52.

KPM, The Ministry of Education Malaysia. (2006). *Syarat dan peraturan menjalankan penyelidikan pendidikan di institusi-institusi di bawah Kementerian Pelajaran Malaysia.* (BPPDPBIL.1/2006) Putrajaya: Federal Government Administrative Center.

KPM, The Ministry of Education Malaysia. (2015). *Pelaksanaan rintis Program Dwibahasa atau Dual Language Program (DLP) di sekolah pada tahun 2016.* (KPMBPSH-SPDK100-1/7/323) Putrajaya: Federal Government Administrative Center.

KPM, The Ministry of Education Malaysia. (2016a). *Pelaksanaan rintis Program Dwibahasa atau Dual Language Program (DLP) di sekolah pada tahun 2017.* (KPMSP.100-6/1/99 Jld.454) Putrajaya: Federal Government Administrative Center.

KPM, The Ministry of Education Malaysia. (2016b). *Dokumen standard kurikulum dan pentaksiran*. Putrajaya: Bahagian Pembangunan Kurikulum.

Krashen, S. D. (1982). Principles and Practice in Second Language Acquisition. Retrieved from www.sdkrashen.com/content/books/principles_and_practice.pdf

Krashen, S. D., & Brown, C. L. (2007). What is academic language proficiency. *STETS Language & Communication Review*, *6*(1), 1-5.

Kremmel, B. & Schmitt, N. (in press). Vocabulary Levels Test. *TESOL Encyclopedia of English Language Teaching*. Wiley-Blackwel.

Lasagabaster, D., (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal, 1*, 30-41.

Lasagabaster, D., & Sierra, J. M. (2010). 'Immersion and CLIL in English: more differences than similarities,' *ELT Journal 64(4),* 367–375.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different?. *Applied linguistics*, *19*(2), 255-271.

Llinares, A. (2015). Integration in CLIL: a proposal to inform research and successful pedagogy. *Language, Culture and Curriculum, 28(1)*, 58-73.

Lorenzo, F., Casal, S., & Moore, P. 2010. 'The effects of content and language integrated learning in European education: key findings from the Andalusian sections evaluation project,' Applied Linguistics 31: 418–42.

Lorenzo, F., & Rodríguez, L. (2014). Onset and expansion of L2 cognitive academic language proficiency in bilingual settings: CALP in CLIL. *System*, *47*, 64-72.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in second language acquisition, 26(3)*, 399-432.

Lyster, R., & Ballinger, S. (2011). Content-based language teaching: Convergent concerns across divergent contexts. *Language Teaching Research*, *15*(3), 279-288.

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Routledge.

Malhotra, N. K. (2006). Questionnaire design and scale development. In R. Grover, M. Vriens (eds.), *The handbook of marketing research: Uses, misuses, and future advances* (pp.176-202). California: Sagepub.

Marsh, D. (Ed.). (2002). CLIL/EMILE: The European dimension. Jyväskylä: University of Jyväskylä. Retrieved from https://jyx.jyu.fi/dspace/bitstream/handle/123456789/47616/david_marsh-report.pdf?sequence=1

Marsh, D., & Frigols, M. J. (2007). CLIL as a catalyst for change in languages education. *Babylonia* 3.07:33-38. Retrieved from http://babylonia.ch/file admin/user_upload/documents/2007-3/marsh_frigols.pdf

Mattheoudakis, M., Alexiou, T., & Laskaridou, C. (2014). To CLIL or not to CLIL? The case of the 3rd Experimental Primary School in Evosmos. In *Selected papers from the 20th International Symposium of Theoretical and Applied Linguistics* (pp. 215-234).

Md Yassin, S., Ong, E. T., Alimon, H., Baharom, S., & Lai, Y. Y. (2010). Teaching Science through English: Engaging pupils cognitively. *International CLIL Research Journal, 1(3)*, 46-59.

Merikivi, R., & Pietilä, P. (2014). Vocabulary in CLIL and in Mainstream Education. *Journal of Language Teaching & Research*, *5*(3).

Met, M. (1991). Learning language through content: Learning content through language. *Foreign Language Annals, 24(4),* 281-295.

Met, M. (1999). Content-based instruction: Defining terms, making decisions. NFLC Reports. Washington, DC: The National Foreign Language Center.

Met, M., & Lorenz. E. B. (1997). Lessons from U.S. immersion programs: Two decades of experience. In R. K. Johnson & M. Swain (eds), *Immersion education: International perspectives* (243-264). Cambridge: Cambridge University Press.

Meyer, O., Coyle, D., Halbach, A., Schuck, K., & Ting, T. (2015). A pluriliteracies approach to content and language integrated learning – mapping learner progressions in knowledge construction and meaning-making. *Language, Culture and Curriculum, 28(1),* 41-57, DOI: 10.1080/07908318. 2014.1000924

Moghadam, N. Z., & Fatemipour, H. (2014). The Effect of CLIL on Vocabulary Development by Iranian Secondary School EFL Learners. *Procedia-Social and Behavioral Sciences*, *98*, 2004-2009.

Mohamad, M. F. A. (2016). Basic statistical analysis: Step by step using SPSS. Johor: UTHM.

Mohd. Saat, R., & Othman, J. (2010). Meeting linguistic challenges in the science classroom: pre-service ESL teachers' strategies. *Asia Pacific Journal of Education*, *30*(2), 185-197.

Muñoz, C. (2007). CLIL: Some thoughts on its psycholinguistic principles. *Revista Española de Lingüística Aplicada, Extra (1)*, 17-26.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English?. *Reading research quarterly*, 304-330.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American educational research journal*, *24*(2), 237-270.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91-108.

Nambiar, R. M. K. (2007) Enhancing academic literacy among tertiary learners: A Malaysian experience. 3L The Southeast Asian Journal of English Language Studies, 13, 77-94.

Nikula, T. (2015). Hands-on taks in CLIL science classrooms at sites for subject-specific language use and learning. *System, 54,* 14-27.

Olsson, E. (2015). Progress in English academic vocabulary use in writing among CLIL and non-CLIL students in Sweden. *Moderna språk*, *109*(2), 51-74.

Olsson, E., & Sylvén, L. K. (2015). Extramual English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Journal of Applied Language Studies, 9(2),* 77-103.

Ong, S. L., & Tan, M. (2008). Mathematics and Science in English: Teachers experience inside the classroom. *Jurnal Pendidik dan Pendidikan, 23,* 141-150. Retrieved from http://apjee.usm.my/APJEE_23_2008/JPP23-8_MATHS%20AND%20SCIENCE%20IN%20ENGLISH.pdf

Paran, A. (2013). Content and language integrated learning: Panacea or policy borrowing myth? Applied Linguistics Review, 4, 317–342. doi:10.1515/applirev-2013-0014

Pearson, N. (2004). The idiosyncrasies of out-of-class language learning: A study of mainland Chinese students studying English at tertiary level in New Zealand. In *Proceedings of the Independent Learning Conference 2003*.

Pérez-Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, *15*(3), 315-341.

Pietilä, P., & Merikivi, R. (2014). The Impact of Free-time Reading on Foreign Language Vocabulary Development. *Journal of Language Teaching & Research*, *5*(1).

Piirainen-Marsh, A., & Tainio, L. (2009). Other-repetition as a resource for participation in the activity of playing a video game. The Modern Language Journal, 93(2), 153–169.https://doi.org/10.1111/j.1540-4781.2009.00853.x

Pladevall-Ballester, E., & Vallbona, A. (2016). CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills. *System*, *58*, 37-48.

Poor English impedes lessons. (2006, December 10). *The Star Online.* Retrieved from https://www.thestar.com.my/news/education/2006/12/10/poor-english-impedes-lessons/

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language learning*, *52*(3), 513-536.

Roessingh, H. (2005). BICS-CALP: An introduction for some, a review for others. *TESL Canada Journal*, *23*(2), 91-96.

Roff, K. A. (2006). Teacher Experiences in Low SES and High SES Schools. *Online Submission*. Retrieved from https://eric.ed.gov/?id=ED493371

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. Studies in Second Language Acquisition, 21, 589–619.

Rott, S. (2012). Incidental Vocabulary Acquisition. *The Encyclopedia of Applied Linguistics.* DOI: 10.1002/9781405198431.wbeal0531

Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: a longitudinal study in the Basque Country. International CLIL Research Journal 1 (1), 60e73.

Ruiz de Zarobe, Y., & Cenoz, J. (2015). Way forward in the twenty-first century in content-based instruction: moving towards integration. *Language, Culture and Curriculum*, *28*(1), 90-96.

Scarcella, R. (2003). Academic English: A Conceptual Framework. Technical Report 2003-1. *University of California Linguistic Minority Research Institute*.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. Linguistics and Education, 12, 431-459.

Schmitt, N. (2000). Introduction. In *Vocabulary in language teaching (pp. 1-9)*. New York: Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. Language Teaching Research, 12, 329–363.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language Testing, 18, 55–88.

Selamat, A., Esa, A., Saad, S. S., & Atim,, A. (2011). Teaching and learning Mathematics and Science in English in primary schools in the state of Johor, Malaysia. *Journal of Education*. 16, 61-73. Retrieved from http://eprints.uthm.edu.my/3825/

Selangor. (n.d.). Retrieved June 15, 2017 from Wikipedia: https://en.wikipedia.org/wiki/Selangor

Singh, P., Rahman, A. A., & Hoon, T. S. (2010). Languages and mathematics achievements among rural and urban primary four pupils: A Malaysian experience. *Journal of Science and Mathematics Education in Southeast Asia*, *33*(1), 65-85.

Snow, M. A., Met, M., & Genesee, F. (1989). A conceptual framework for the integration of language and content in second/foreign language instruction. *Tesol Quarterly*, *23*(2), 201-217.

Sundqvist, P. (2009). The impact of spare time activities on students' English language skills. In *CSL 2009*, 63-76. Retrieved from http://www.diva-portal.org/smash/get/diva2:292259/FULLTEXT01.pdf

Swain M., (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S.M. Gass, C.G. Madden (eds.), *Input in Second Language Acquisition (*235- 253). Rowley, MA: Newbury House Publishers, Inc.

Swain, M. (1988). Manipulating and complementing content teaching to maximize second language learning. *TESL Canada Journal, 6(1),* 68-83.

Sylvén, L. K. (2013). CLIL in Sweden–Why does it not work? A metaperspective on CLIL across contexts in Europe. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 301-320.

Sylvén, L. K. (2004). *Teaching in English or English teaching? On the effects of content and language integrated learning on Swedish learners' incidental vocabulary acquisition* (Doctoral dissertation, Göteborg University).

Taboada, A., & Rutherford, V. (2011). Developing reading comprehension and academic vocabulary for English language learners through science content: A formative experiment. *Reading Psychology*, *32*(2), 113-157.

Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research*, 15(3), 325-342. doi: 10.1177/1362168811401153

Tedick, D. J., & Cammarata, L. (2012). Content and language integration in K–12 contexts: Student outcomes, teacher practices, and stakeholder perspectives. *Foreign Language Annals*, *45*(1), 28-52

Tedick, D. J., Christian, D., & Fortune, T. W. (2011). The future of immersion education: An invitation to 'dwell in possibility'. In D. J. Tedick, D. Christian & T. W. Fortune, *Immersion education: Practices, policies, possibilities* (1-10). Bristol: Short Run Press Ltd.

Tedick, D. J., & Wesely, P. M. (2015). A review of research on content-based foreign/second language education in US K-12 contexts. *Language, Culture and Curriculum*, *28*(1), 25-40.

Tragant, E., Marsol, A., Serrano, R., & Llanes, À. (2016). Vocabulary learning at primary school: a comparison of EFL and CLIL. *International Journal of Bilingual Education and Bilingualism*, *19*(5), 579-591.

Várkuti, A. (2010). Linguistic benefits of the CLIL approach: Measuring linguistic competences. *International CLIL Research Journal*, *1*(3), 67-79.

Walker, C. L., & Tedick, D. J. (2000). The complexity of immersion education: Teachers address the issues. *The Modern Language Journal, 84(1),* 5-27. https://doi.org/10.1111/0026-7902.00049

Webb, S. A., & Chang, A. C. S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113-126.

Wiesemes, R. (2009). Developing theories of practices in CLIL: CLIL as post-method pedagogies? In Y. Ruiz de Zarobe & R. M. J. Catalán (Eds), *Content and language integrated learning. Evidence from research in Europe* (41-63). Bristol: Short Run Press Ltd.

Xanthou, M. (2010). Current trends in L2 vocabulary learning and instruction: Is CLIL the right approach. *Advances in Research on Language Acquisition and Teaching: Selected Papers, Thessaloniki, Greece: Greek Applied Linguistics Association (GALA)*, 459-471.

Xanthou, M. (2011). The impact of CLIL on L2 vocabulary development and content knowledge. *English Teaching*, *10*(4), 116.

Zulkefli, A. S. (2018, January 5). 126 more schools approved for Dual Language Programme. *New Straits Times.* Retrieved from https://www.nst.com.my/news/nation/2018/01/321822/126-more-schools-approved-dual-language-programme