

**Project Title:** Informing pedagogy through exploring and reforming  
assessment practices in EMI education

**Grantee:** The University of Hong Kong

**Principal** LO Yuen-yi

**Investigator:** Faculty of Education  
The University of Hong Kong

**Co-** David CARLESS

**investigators:** Faculty of Education  
The University of Hong Kong

FUNG Chun-lok, Dennis  
Faculty of Education  
The University of Hong Kong

Angel M.Y. LIN  
Faculty of Education  
The University of Hong Kong

# Final Report

by

Principal Investigator

# **Research project: Informing pedagogy through exploring and reforming assessment practices in EMI education**

## **Final report**

### **(a) Abstract**

To facilitate second/foreign language learning, it has become more popular to use the target language as the medium of instruction of non-language content subjects. This trend is widely recognised as Content and Language Integrated Learning (CLIL) and the English-as-medium-of-instruction (EMI) education in Hong Kong can be regarded as one of its variants. To date, very limited research has examined how students are assessed and whether assessment practices align with the dual goal of CLIL (i.e. content and language learning) and classroom teaching. This three-phase study sought to address these important questions. The first stage examined the questions in different types of assessment, and revealed a rather big leap in both cognitive and linguistic demands from junior to senior secondary education. The second phase of the study examined the relationship among objectives, instruction and assessment practices of 12 Biology/Integrated Science and Geography teachers, employing a multi-case study approach. It was observed that only a few teachers incorporated explicit language scaffolding to help students meet the linguistic demands of assessments. In the final stage, an assessment paper was designed for Science and Geography according to this study's theoretical framework. It was tried out with students, whose performance was analysed to see if the tests could diagnose students' learning in content and language dimensions. The findings of this study deepen our understanding of assessment practices in EMI and yield important implications for designing valid assessments in EMI. These also inform more effective classroom pedagogy and enhance the learning effectiveness of EMI/CLIL education.

**(b) Keyword(s):** Content and Language Integrated Learning (CLIL);

### **(c) Introduction & Background**

In Hong Kong, there has always been a strong demand for high levels of English (L2) proficiency, which can be attributed to a combination of historical, political and socio-economic factors (Tsui, 2004). Consequently, there has been an overwhelming preference for English as the medium of instruction (EMI) in secondary schools (Choi, 2003), which are believed to facilitate English learning. In these EMI schools, students learn most non-language content subjects (e.g. mathematics, science, history) in English (L2) and sit for internal and external examinations in English.

Although the EMI education in Hong Kong has its historical root, it actually coincides with the increasingly popular trend of using students' target language as the medium of instruction in non-language content subjects in other parts of the world. Such kind of programmes is generally grouped under the umbrella term Content and Language Integrated Learning (hereafter CLIL<sup>1</sup>) (Cenoz et al., 2014). The rationale behind such programmes is that content subjects provide authentic communicative contexts for students to be exposed to more input, interaction and output opportunities, which are favourable for L2 learning (Lyster & Ruiz de Zarobe, 2017).

The worldwide spread of CLIL to different educational contexts, especially to those English as a foreign language (EFL) ones, has attracted a great deal of research efforts, which to date have largely focused on student achievements and classroom discourse (Pérez-Cañado 2012), leaving assessment 'a blind spot' in many CLIL programmes (Massler et al., 2014, 138). This issue deserves urgent attention for three reasons. First, it has been observed that students

---

<sup>1</sup> As EMI can be regarded as a variant of CLIL, EMI and CLIL are used interchangeably in this report.

express their content knowledge better in their L1 (Gablasova, 2014) and they tend to perform high-order thinking skills in their L1 (Luk & Lin, 2015). Hence, CLIL students being assessed of content knowledge through their less proficient L2 raises the concern about the validity of CLIL assessments. Second, language (i.e. the target language in CLIL) should actually be assessed, given the dual-focus on content and language learning in CLIL. However, how to design valid assessments that diagnose student learning progress (and also difficulties) in both the content and language dimensions is complex (Heine, 2014). CLIL teachers often think they assess content knowledge only, but they actually assess both during the marking process (Hönig, 2010). Third, it has been argued that assessment has ‘backwash effect’ on teaching and learning behaviour (Alderson & Wall, 1993). Hence, delving into the assessment issues in CLIL can inform what teachers and learners have to do in order to move towards their learning targets in terms of content and language.

This study seeks to contribute to the under-researched area of CLIL assessments by exploring the current assessment practices of CLIL, examining the alignment among objectives, instruction and assessment, and designing assessments that may better diagnose students’ learning progress in both content and language dimensions. By investigating CLIL assessments from different perspectives, this study will provide important insights for policy makers and teaching practitioners into designing valid assessments, and into supporting students to tackle assessments of content subjects in CLIL.

#### **(d) Review of literature of the project**

##### ***Growing research on CLIL/ EMI and the gap in assessment***

The benefits of CLIL in other contexts have been well documented. For example, in the Canadian immersion programmes, research has shown that immersion students outperformed

their non-immersion peers in L2 (French) proficiency, with no detriment to their academic achievements (see a recent review by Lazaruk, 2007). In the European context, a growing body of research has pointed to similar benefits that CLIL students enjoy in terms of L2 learning (see Pérez-Cañado's review, 2012), but their academic achievements have not been widely examined (Cenoz et al., 2014). In Hong Kong, empirical studies on EMI education over the past decades have demonstrated that EMI students enjoyed some advantages in L2 (English) learning, yet they were achieved at the expense of their achievement in such content subjects as science and history (e.g. Marsh et al., 2002; Education Bureau, 2006; see also Lo & Lo's meta-analysis, 2014).

Therefore, it seems that the dual goal of content and language learning in CLIL/ EMI is not guaranteed. To explain for the inconsistencies across programmes and contexts and to enhance the effectiveness of CLIL, recent research has paid more attention to classroom interaction and discourse (Nikula et al., 2013), so as to examine the teaching and learning processes in CLIL lessons, particularly how teachers and students co-construct content and language at the same time (e.g. Lin & Wu, 2015).

Despite a great deal of research effort and attention to the increasingly popular CLIL, there has been an underexplored area – assessment (Hönig, 2010; Massler et al., 2014). The assessment issues in CLIL are not only important, as aforementioned, but also highly complicated. They will be illustrated in detail below.

### ***Issues with assessments in CLIL/ EMI***

The role of assessment in previous CLIL research has largely been the instrument to evaluate the effectiveness of CLIL programmes, so that researchers could compare the achievement of

CLIL and non-CLIL students in their L2 proficiency and content subject knowledge. However, very few researchers have questioned the validity of assessment used in CLIL. In educational assessment and testing literature, validity concerns whether the test score can accurately reflect a student's level of knowledge, skills or competencies which the test is intended to measure (Hughes, 2003; Shaw & Imam, 2013). In this way, the test score can be appropriately interpreted and used (Kane, 2006). Applying the concept of assessment validity to the CLIL context, it refers to whether assessment in CLIL measures what it targets at and whether assessment can reasonably reflect students' actual learning. This interpretation may look straightforward, but is actually very complicated if one considers the dual goal of CLIL and students are assessed for both L2 competencies and knowledge in content subjects.

When assessing students' knowledge and skills involved in non-language content subjects in CLIL, the validity issue deserves serious attention as students are assessed through their less proficient L2. It has been shown that students could better express their content knowledge in their first language (L1), when compared to their performance on the same task in L2 (Gablasova, 2014). Therefore, assessment in CLIL may bear the risk of not accurately reflecting (very likely underestimating) students' actual knowledge in content subjects. Through text analysis of the internationally recognised IGCSE examination papers (including both the questions and instructions) as well as students' scripts, Shaw (2012) and Shaw and Imam (2013) and identified the linguistic demands that various subjects (biology, geography and history) imposed on candidates and they further evaluated the threshold English level that candidates needed to access those examinations. The researchers observed that candidates' low scores in the examinations were mainly the result of deficiencies in their subject knowledge rather than linguistic hindrance. Yet, the researchers did point out that candidates with insufficient linguistic resources may not achieve the maximum marks on questions

requiring more developed answers (e.g. essay type questions in history and geography). In other words, to attempt examinations of content subjects in an L2, students do need to possess a certain level of academic language, especially those subject-specific vocabulary and certain general academic vocabulary (e.g. identify, evaluate, state) so that they understand what the tasks or questions require. To succeed or excel in those examinations, students further need to be equipped with more linguistic resources to organise and present their ideas in a better way. These studies have yielded important implications for the role of language in assessment in CLIL, especially in contexts where the target language is students' foreign language and where students are taking high-stakes examination in the target language (e.g. in Hong Kong).

The validity of assessment in CLIL is further complicated by the dual goal of the programme. Both Short (1993) and Coyle et al. (2010) have argued, one core issue regarding assessment in CLIL is "*what to assess?*", in particular, whether the focus should be on content or language, or on both. Theoretically speaking, both content and language should be assessed as they are the dual goals in CLIL (Massler et al., 2014). Practically, CLIL content subject teachers, on one hand, do not think they target at both when they design the assessment tasks and marking rubrics (Massler et al., 2014), but they, on the other hand, are actually examining both content and language implicitly as students have to understand the assessment questions and express their content knowledge through language. In Hönig's study (2010), in which the CLIL History teachers stated in the interviews that they would only consider the content knowledge that students expressed in their oral presentation, yet when they marked the oral presentations, students' oral proficiency did play an important role in those teachers' grading, in the sense that the teachers justified their grading with reference to students' language proficiency.



### *Alignment between objectives, instruction and assessment*

Another issue related to the validity of assessment is classroom practices. It has been suggested that valid assessment should align with programme/ lesson objectives as well as the instruction in classrooms (Orlich et al., 2013). If students are not assessed of what they have been taught, the assessment does not serve the purpose of indicating students' progress and the effectiveness of teaching. In many CLIL educational contexts, content subject lessons are taught by content subject specialists (Mehisto, 2008). Hence, content subject teachers in CLIL have been observed to put more emphasis on teaching content (Walker, 2011; Tan, 2011), probably due to their lack of language awareness (Lo, 2014a; Trent, 2010) and/or lack of language teaching pedagogy (Koopman et al., 2014). If that is the case, it would not be valid to assess both students' content knowledge and L2 proficiency in examinations. On the other hand, assessment has an impact on classroom practices, which is commonly known as the "backwash" effect (Alderson & Wall, 1993) and has been widely investigated in language learning classrooms. In the CLIL context, when high-stakes examinations of content subjects (e.g. the Hong Kong Diploma of Secondary Education Examination, HKDSE) put more emphasis on content knowledge, the backwash effect could be that CLIL subject teachers do not see the need to scaffold students' academic literacy development or to incorporate more language teaching in their lessons. Hence, it would be worth investigating the intriguing relationship between objectives, classroom practices and assessment, especially how they may affect each other.

The above literature review of CLIL assessment has identified two core issues. The first one is how assessment in CLIL content subjects can evaluate students' content and language learning outcomes in a valid way. In other words, it concerns whether assessment in CLIL can diagnose students' learning progress (or difficulties) in both content and language dimensions. The

second issue is the relationship between assessment and classroom practices in CLIL classrooms. From these two core issues, the following research questions are formulated:

1. How valid are current assessment practices in CLIL, in terms of assessing students' content and language learning?
2. To what extent does assessment affect classroom practices or vice versa?
3. How can assessment tasks be designed so as to promote content and language integrated learning?

#### **(e) Theoretical and/or conceptual framework of the study**

Without a clear framework guiding the design of assessment in CLIL, the assessment tasks designed may not be valid in the sense that they may put too much emphasis on content knowledge, instead of an integration of content AND language. In addition, the assessment tasks may not be able to diagnose whether students have grasped the target concepts, or students are hindered by language barriers, or both (to varying extents). Lo and Lin (2014) put forward a theoretical framework for teachers to analyse the linguistic and cognitive demands that different assessment tasks impose on students in CLIL. In this framework, assessment tasks are evaluated by their “cognitive demand”, which can be divided into three levels, namely “recall”, “application” and “analysis”. These levels are adapted from the six levels in Bloom’s taxonomy (Bloom, 1956; Krathwohl, 2002). Along the other dimension are three levels of “linguistic demand”, which include “vocabulary”, “sentence patterns” and “text”. These three levels correspond to the various features of academic language identified (Schleppegrell, 2004). The framework thus provides a useful analytical tool for this study to examine the validity of assessment practices in relation to their cognitive and linguistic demands.

After applying the original Lo & Lin's framework (2014) to the data collected for this study, some modifications were made so as to analyse the data more accurately (see the revised framework in Figure 1). First, the linguistic demand was separated into "receptive" (i.e. reading the question) and "productive" (i.e. writing an answer) demands. This is because every question contains receptive linguistic information to decode, the level of which could be different from the productive language requirement. For example, a typical essay-type question exerts a receptive *sentence* demand and a productive *text* demand because students need to read the questions presented in sentences, and write their answer in an essay.

Second, the revised framework contains an additional level of "no productive linguistic demand". This level is typically represented by multiple-choice questions (as students are simply asked to write the letters corresponding to the answers they choose), graph plotting and calculations. The revised framework is believed to be able to generate a more specific and fine-grained analysis of assessment questions.

|                   |                      |                                 | Cognitive demand |             |          |
|-------------------|----------------------|---------------------------------|------------------|-------------|----------|
|                   |                      |                                 | Recall           | Application | Analysis |
| Linguistic demand | Receptive Vocabulary | No productive linguistic demand |                  |             |          |
|                   |                      | Productive vocabulary           |                  |             |          |
|                   |                      | Productive sentence             |                  |             |          |
|                   |                      | Productive text                 |                  |             |          |
|                   | Receptive sentence   | No productive linguistic demand |                  |             |          |
|                   |                      | Productive vocabulary           |                  |             |          |
|                   |                      | Productive sentence             |                  |             |          |
|                   |                      | Productive text                 |                  |             |          |
|                   | Receptive text       | No productive linguistic demand |                  |             |          |
|                   |                      | Productive vocabulary           |                  |             |          |
|                   |                      | Productive sentence             |                  |             |          |
|                   |                      | Productive text                 |                  |             |          |

Figure 1. The framework adopted in this study to analysis the linguistic\content demand of assessment tasks

#### (f) Methodology & Data collection and analysis

A three-phase study was conducted in the EMI education in Hong Kong to address the research questions.

##### *Stage 1: Analysis of current assessment practices*

The aim of this stage was to survey the current assessment practices in the EMI education in

Hong Kong. This stage involved analysis of a collection of assessment tasks used in EMI schools in Hong Kong. Among the wide range of content subjects in EMI education, this study focused on two, namely Integrated Science (junior secondary)/ Biology (senior secondary) and Geography. The choice of these two subjects was justified by two reasons. First, it is desirable to include one Science and one Humanities subject, as it has been observed that the linguistic demands involved in different subject disciplines tend to be different (Lo, 2014b). Second, the two particular subjects are selected because they are offered by over 90% of secondary schools in Hong Kong (HKEAA, 2013) and they are found across different stages in the secondary school curriculum (including both junior and senior levels).

#### *Data collection*

The data of this stage were mainly assessment questions in EMI and they came from three sources:

- (i) Questions were collected from a set of Science, Biology and Geography textbooks and the accompanying workbooks, which was selected based on its popularity among local secondary schools. These textbooks were produced by the same publisher. There were altogether 2491 questions from junior form Science, 1940 questions from senior form Biology, 1386 questions from junior form Geography, and 500 questions from senior form Geography. These represent continuous and formative assessments in schools.
- (ii) The end-of-term/year examination papers set by Biology/ Integrated Science and Geography teachers in 10 local EMI secondary schools were also gathered. There were 767 Science/Biology and 807 Geography questions. These represent the school-based summative assessment practices designed by content subject

teachers.

- (iii) Questions were gathered from the annually held HKDSE from 2012 to 2015.

There were 387 Biology and 354 Geography questions. These represent high-stakes summative assessments.

### *Data analysis*

The assessment questions sampled were coded using the framework shown in Figure 1. The unit of analysis was each question or each part of the multi-part assessment question. This is considered a more appropriate unit of analysis because questions are sometimes broken down into several parts and each part might have a distinctive demand different from the other parts of the same question. The coding of assessment questions was conducted separately by two research team members to ensure a satisfactory inter-rater reliability. Any discrepancies were discussed and resolved by the coders. The distribution of the cognitive and linguistic demands of the questions found in different types of assessments was then summarised and compared to address research question 1.

### ***Stage 2: Alignment among objectives, instruction and assessment***

This second stage of this study employed a multiple case-study approach, so as to have a holistic and in-depth investigation of how objectives, classroom practices and assessment may interact or affect each other in particular school contexts. One content subject teacher teaching Biology/ Integrated Science or Geography through EMI constituted one case. Invitation letters were sent to over 100 secondary schools in Hong Kong, and eventually 12 teachers, 5 teaching Biology/ Integrated Science and 7 Geography, were recruited. These teachers came from 9 schools in different districts and with students of different socio-economic status and academic abilities. They also had different years of teaching

experience and professional training. Hence, a comparison across cases may yield more interesting insights into the research question.

#### *Data collection*

Several sources of data were collected from each case teacher to understand his/her objectives, classroom practices and assessment practices.

- (i) Lesson observations: Each teacher was observed when teaching one unit or topic of the subject (around 3-6 lessons). These lessons were video or audio-recorded and at least one research team member was present in the classroom to jot field notes. Whenever possible, the researcher would have a brief chat with the teacher before and after the lesson observation to understand his/her lesson plan and objectives and classroom practices.
- (ii) Collection of assessment tasks: The formative and summative assessment tasks (i.e. including homework, end-of-unit quizzes, end-of-term/year examinations), together with the marking rubrics, used by each case teacher for the unit/topic observed were collected. In addition, a stratified sample of marked scripts were provided by each teacher so as to allow the research team to analyse the students' performance and the teachers' marking practices.
- (iii) Semi-structured interviews: The perceptions of both teachers and students of the classroom practices and assessment in EMI schools were collected, so as to triangulate or elaborate on what was observed in lessons and analysis of assessment tasks. An individual semi-structured interview was conducted with each case teacher, whereas focus group interviews were conducted with one group of students (3-4 students in a group) from the class where the lessons were observed.

### *Data analysis*

Relevant parts of the observed lessons and interviews with teachers and students were transcribed verbatim to allow for detailed analysis. These, together with the documents collected, were analysed and coded for recurrent themes or categories related to the research question (i.e. the alignment among objectives, instruction and assessment practices).

In particular, the transcribed lessons were analysed according to their functions and foci. First, teacher and student utterances were classified into instructional or regulative register. Second, for those under instructional register, which would be the key focus of this study, they were further classified according to their focus on “content” or “language”, and then their respective level (e.g. “recall”, “application” and “analysis” for “content”; “lexico-grammar”, “sentence” and “text” for “language”). The analysis generated by such coding procedures enables the research team to examine the attention paid to content and language in the lessons observed.

Then, the assessment practices were analysed. There were three major types of assessments associated with the observed lessons. The first type was oral formative assessments, which were typically conducted in the form of teachers’ questions in the lessons. These questions were again categorised according to their focus on “content” and “language”. In very few cases, the students were asked to do oral presentation in class, which was also regarded as oral formative assessments. The second type was written formative assessments, which included worksheets and workbook exercises given by the teachers as homework. The third type was written summative assessments, which included formal quizzes, tests or examinations. However, due to the time lag between the observed lessons and the test or examination period, the third type of assessments was not collected from most of the cases.



From the collected assessment tasks, students' performance and teachers' grading practices shown in the marked scripts were analysed.

### ***Stage 3: Trial of content and language integrated assessment tasks***

The third stage attempted to design and try out assessment tasks which can better measure students' content knowledge, L2 proficiency and integration of both.

Four S.2 classes (119 students) from one EMI secondary school participated in this phase of study. The research team and their Integrated Science (IS) and Geography teacher designed an informal test for one particular unit for each subject (test papers attached in Appendix 1). Two classes of students took the IS test and the other two took the Geography test.

As the test papers aimed to measure students' content knowledge, L2 proficiency and integration of both, they were designed with reference to this study's theoretical framework (Figure 1), and included different types of questions varying in levels of cognitive and linguistic demands. For instance, some questions imposed high cognitive demand but low linguistic demand (e.g. multiple-choice questions which mainly asked students to read some statements but required analytical skills); some questions imposed high linguistic demands but low cognitive demands (e.g. students were asked to describe and elaborate on the given impact of technological innovations); and some may be challenging in both (e.g. students were required to write a short coherent paragraph to describe and explain the results of an experiment set up). In this way, the performance of students on a continuum of cognitive and linguistic demands was gathered for analysis. However, it should be noted that considering students' grade level (S.2), the time limit allowed for the test (30 minutes) and also the topic concerned, it was very difficult to include all the possible combination of cognitive and linguistic demands. Yet, the research team tried to maximise the different types of questions.

After finishing the tests, two groups of students (5-6 students in each group) were invited to attend a group interview with the researchers. One group included the students taking the IS test while the other group took the Geography test. The interviews were intended to adopt stimulated recall techniques, in which the students were prompted to recall their thinking processes when they were taking the test, with the test papers as the stimuli. However, it has to be admitted that such techniques were not very successful in tapping into the students' thinking processes, as most students provided very general comments about whether they thought the questions were difficult or not. This is probably due to the fact that the students, aged between 13 and 14, were not used to reporting their thinking processes, which is actually one potential limitation of using stimulated recalls with young research participants.

#### *Data analysis*

The test papers were marked by research assistants who were pre-service teachers of Science or Geography, based on the marking scheme devised by the research team and the teachers in the school. The marks obtained by each individual student in each question or each part of a multiple-part question (e.g. structured questions) were recorded. Then, the marks obtained by an individual student in attempting a specific type of questions (e.g. *recall* questions demanding no production) in a test paper was aggregated and then divided by the number of questions to give a percentage. For example, if a student received 8 marks out of a possible total of 21 marks in all the *recall* questions with no productive language demand, s/he was regarded as scoring 38% in this type of questions. These percentages were then averaged to give the mean percentages of students' performance in a particular type of question. Next, the mean percentages attained in the different types of questions were compared using inferential statistical tests. In this way, how students performed on different levels of cognitive and linguistic demands could be investigated. The interviews with students were transcribed

verbatim and the transcripts were analysed to see if their comments could elaborate or complement the statistical results.

## (g) Results and Discussion

### *Results of Stage 1: Analysis of current assessment practices*

#### *1. Summary of analysis of textbooks and workbooks*

In this section, Science and Biology questions are first presented, followed by Geography.

Figure 2 illustrates the distribution of Science questions.

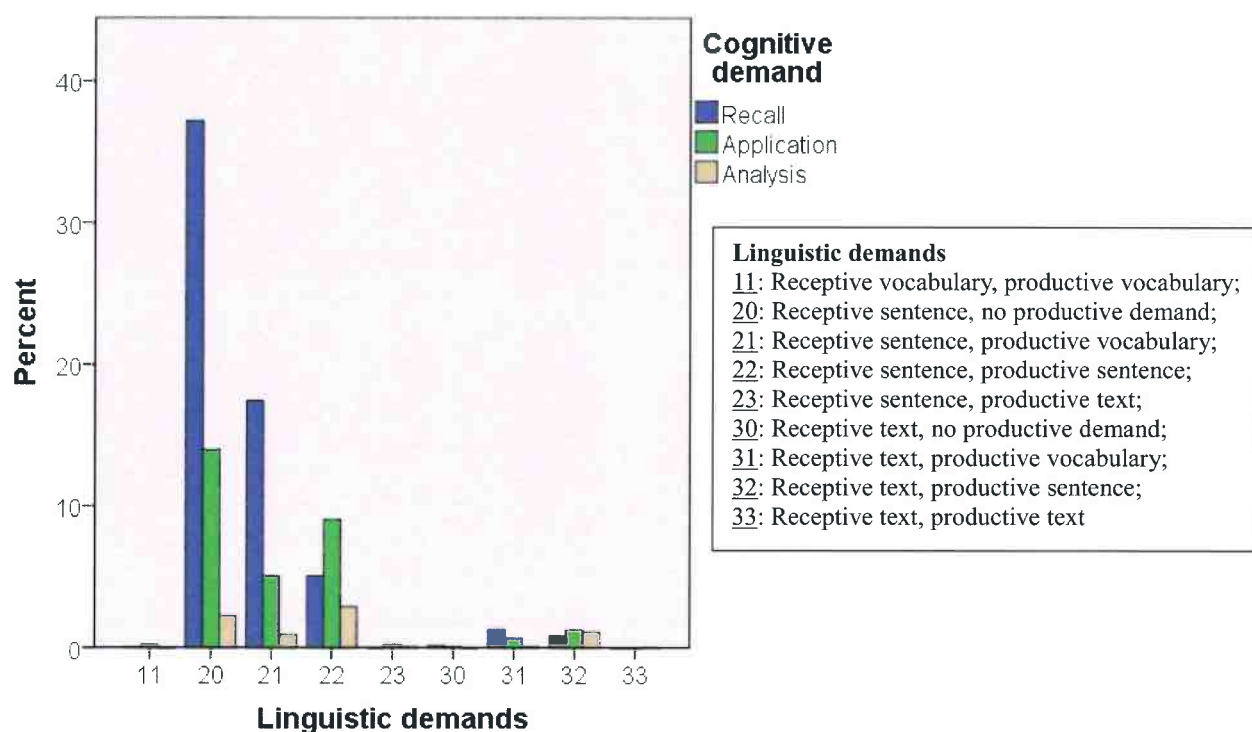


Figure 2. Bar chart showing distribution of Science questions in textbooks/workbooks

It was found that junior secondary Science textbooks/workbooks consisted mainly of recall questions (62.02%), followed by around one-third of application questions. There were only around 7% questions which involved analytical and other higher-order thinking skills. For linguistic demands, the majority of questions required comprehension at the sentence level

(around 92%). Among these questions, more than 75% of the questions demanded no language production or only production at the vocabulary level. Focusing solely on the productive linguistic demands, some sentence production questions existed (around 20% in total), but questions which needed production at the text level was scarce (less than 1%). Taken together, junior form Science questions presented in textbooks/workbooks concentrated on relatively low cognitive and productive language demands.

Figure 3 shows the distribution of senior secondary Biology questions. There were obvious differences when compared to junior form Science. First, there were only 27.27% recall questions, but more than half (53.30%) application questions and 19.43% analysis questions. This suggests that senior form Biology questions were more cognitively demanding. Turning to linguistic demands, while most questions were still presented to students on the sentence level (around 82%), there were considerably more questions presented in form of texts (around 18%). Productive linguistic demands were also moving from the minimal demand in Science to sentence (around 54%) or even text (around 5%) production in Biology. In sum, senior form Biology questions involved relatively higher-order thinking and more advanced language reception and production.

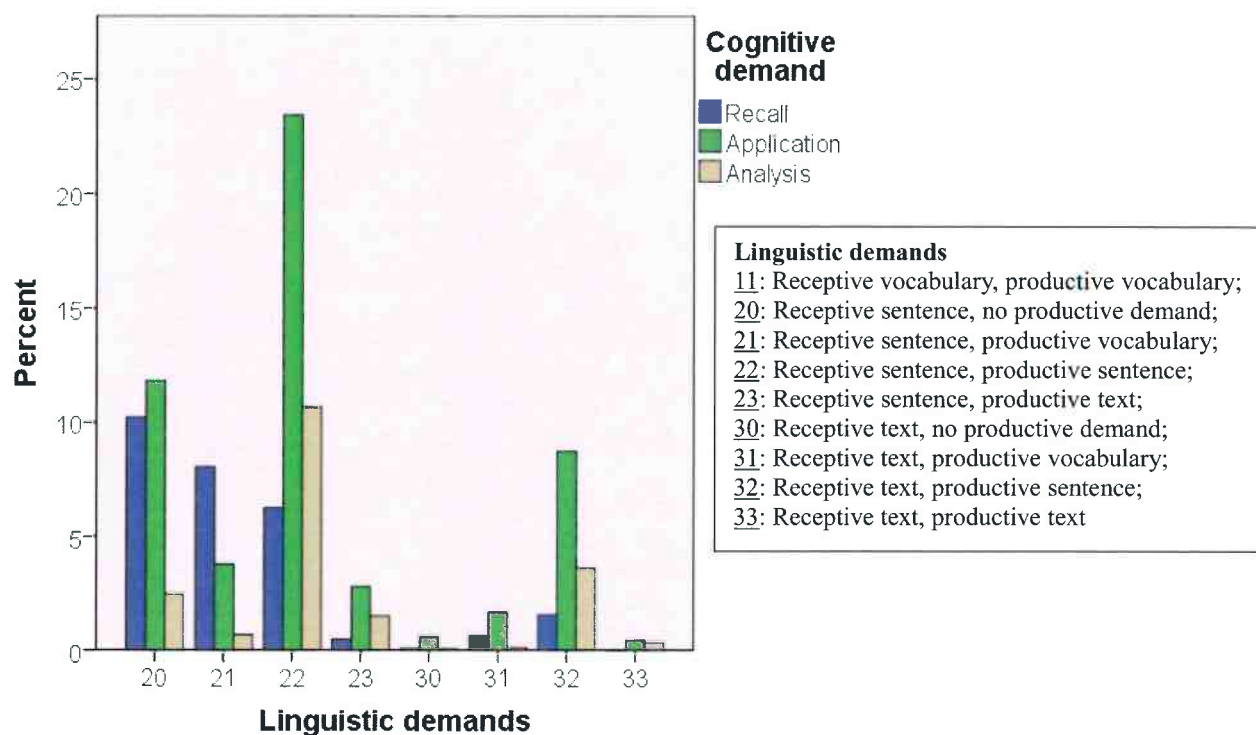


Figure 3. Bar chart showing distribution of Biology questions in textbooks/workbooks

For Geography questions, similar to Science and Biology, it would be interesting to identify any progression of questions in terms of cognitive and linguistic demands. Therefore, the Geography questions used in junior and senior forms textbooks/workbooks are separately analysed. Figure 4 shows the distribution of junior form Geography questions.

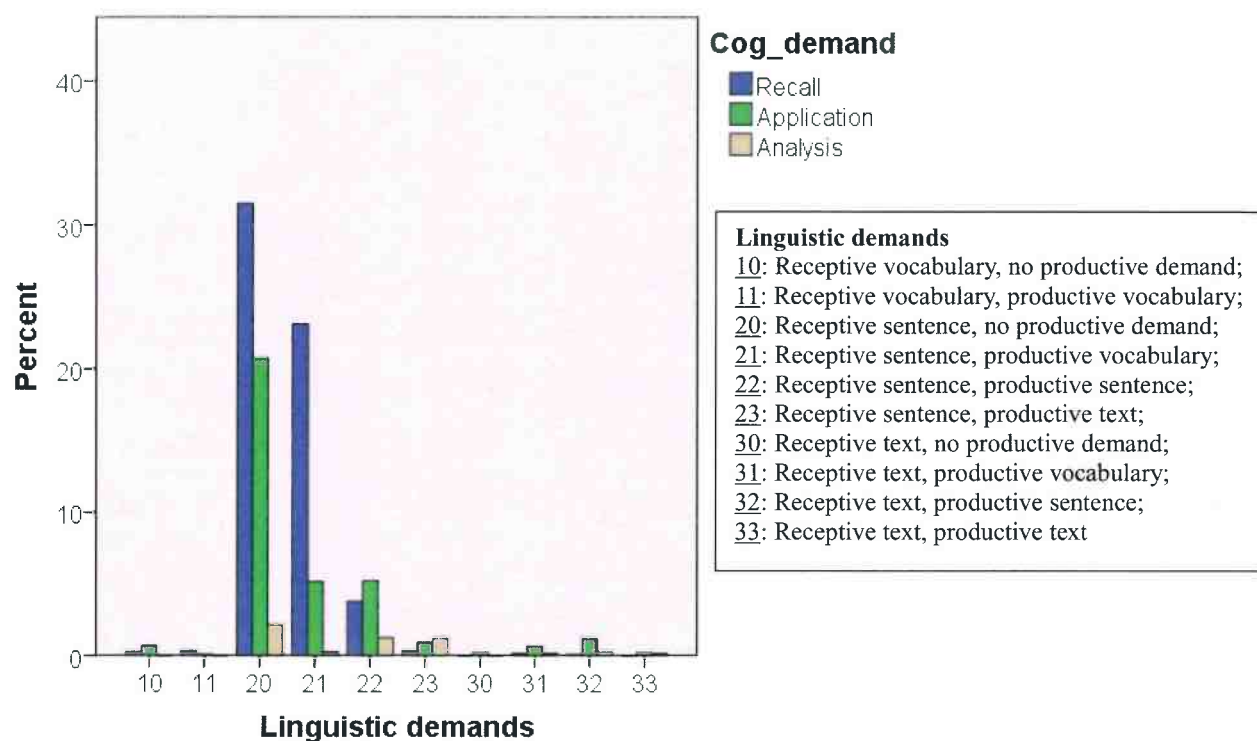


Figure 4. Bar chart showing distribution of junior Geography questions in textbooks/workbooks

It was revealed that junior form Geography questions resembled the trends in Science to a large extent. First, students were mostly tested on their recall of knowledge in almost 60% of the questions. Some 35% of the questions demanded application but there were only 5% of questions necessitating analysis. For linguistic demands, questions presented at the sentence level were dominating (around 95%), and more than 80% of questions required no production or vocabulary production only. Examining specifically the productive linguistic demand, only around 12% required sentence production and 3% text production. It can be concluded, therefore, that junior form Geography questions mirrored those in junior form Science, concentrating on relatively low cognitive and productive language demands.

Figure 5 depicts the distribution of senior form Geography questions. In terms of cognitive demand, there were only 17.80% recall questions, but more than half (52.40%) application questions and considerably more analysis questions (29.80%). Consequently, it can be argued that the senior form Geography questions were more demanding when compared to those in junior forms. For linguistic demand, while almost all questions were still presented at the sentence level, the percentage of questions demanding minimal production (no or vocabulary production) dropped from more than 80% in junior form to less than 60% in senior form. Instead, there were around 30% questions demanding the production of sentences, and 11% that of texts.

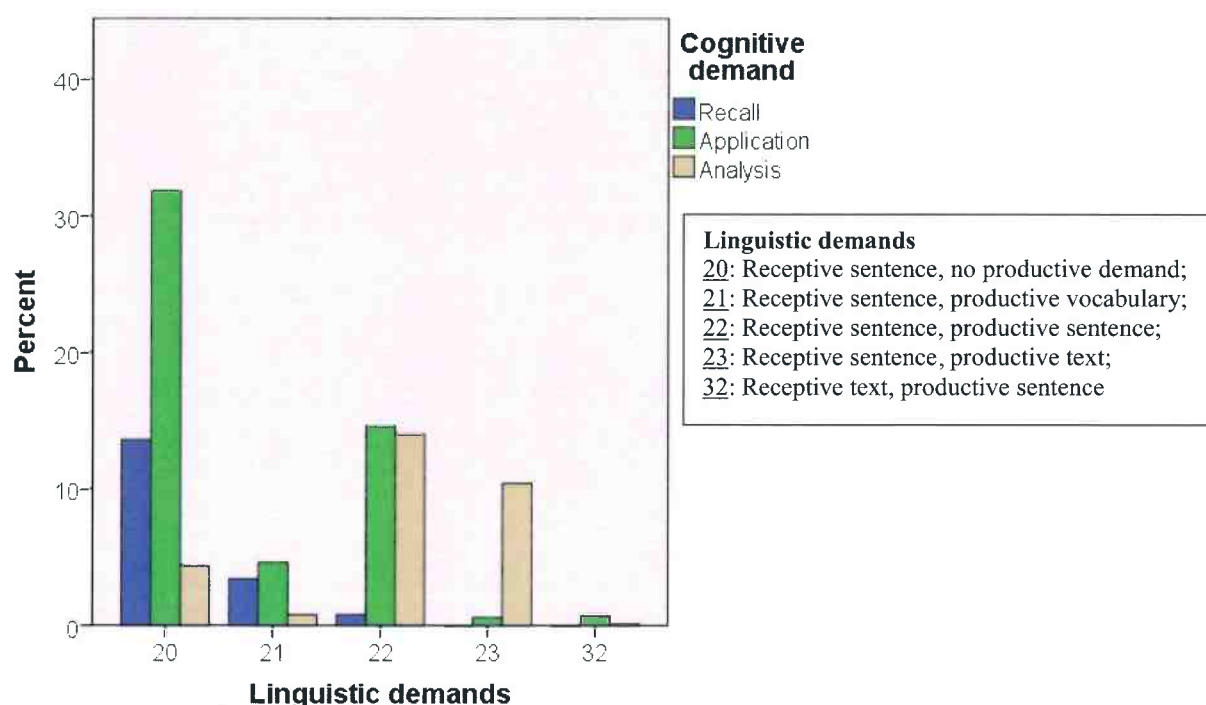


Figure 5. Bar chart showing distribution of senior Geography questions in textbooks/workbooks

## 2. Summary of analysis of school exam papers

The second source of assessment questions analysed in this study came from school examination papers. Figure 6 reveals the analysis of the Science/Biology exam papers.

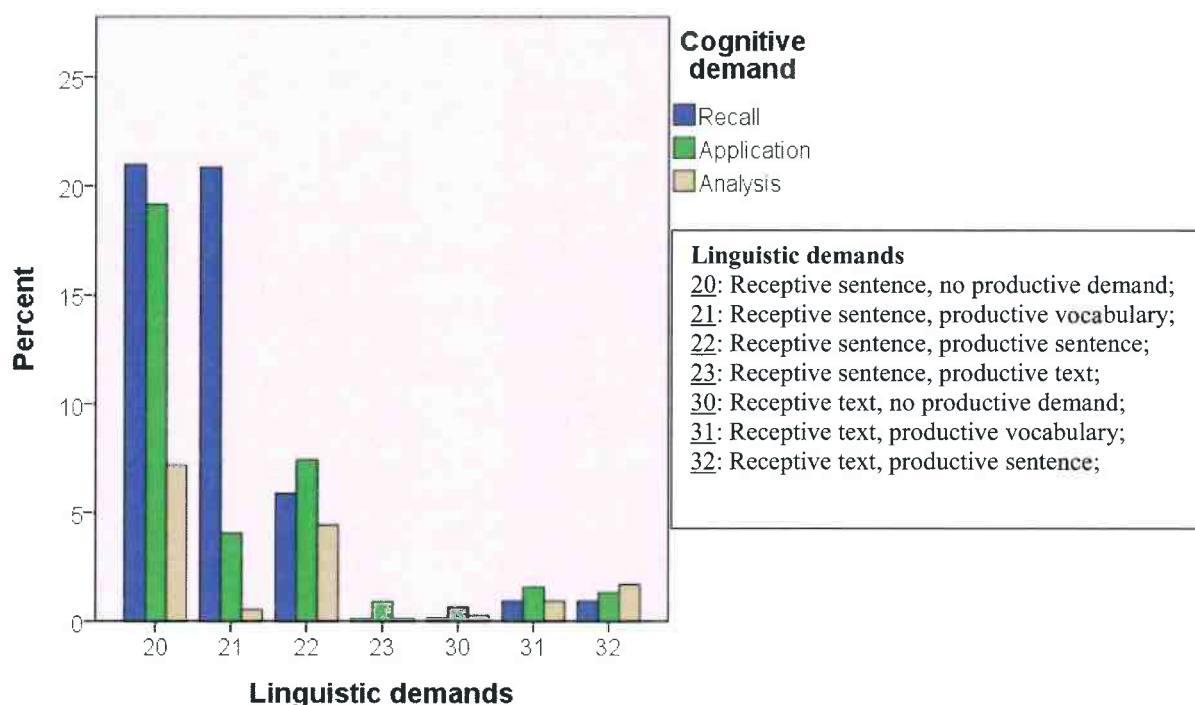


Figure 6. Bar chart showing distribution of Science/Biology questions in school exam papers

For Science/Biology, focusing on the cognitive demand per se, half of the questions required recall of knowledge, and one-third of them application of knowledge. Only around 15% involved analysis of knowledge. Turning to the linguistic demands, most of the questions required understanding of the questions on the sentence level (around 92% in total). Among these questions, the majority required no production (47.33% out of all questions) or production only at the vocabulary level (25.42%). There were not many questions necessitating production at the sentence or text level. In other words, most of the questions involved only relatively low level cognitive processing, as well as low level productive linguistic demand.

Figure 7 shows the analysis of Geography exam papers. First, more than 40% of the questions involved recall and application of knowledge respectively. Similar to Science, only



around 15% of the questions required analysis of knowledge. Linguistically, again there was a predominant proportion of questions possessing a receptive sentence demand (around 92%). Among these questions, the distribution was also very similar to Science, with 46.47% and 26.39% demanding no production and vocabulary production respectively. Although still almost negligible, there were relatively more text production questions for Geography (around 6%). Taken together, both Geography and Science shared the general trend of having more than half of the questions requiring relatively low level cognitive and productive linguistic demand. These align with the trends observed in textbooks. In other words, for junior secondary levels, students encounter similar cognitive and linguistic demands in formative assessments (typically represented by the questions in textbooks and workbooks) and in summative assessments.

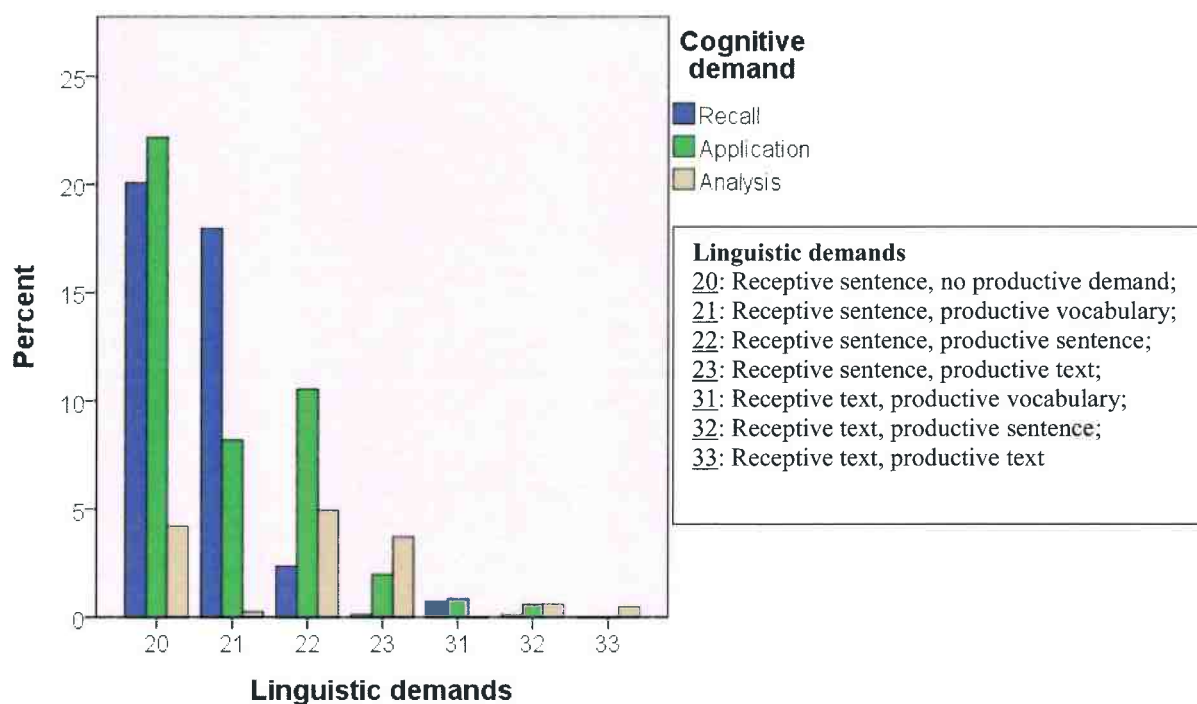


Figure 7. Bar chart showing distribution of Geography questions in school exam papers

### 3. Summary of analysis of HKDSE papers

The last source of assessment questions came from Biology and Geography HKDSE papers from 2012 to 2015. Figure 8 below depicts the analysis of the Biology papers.

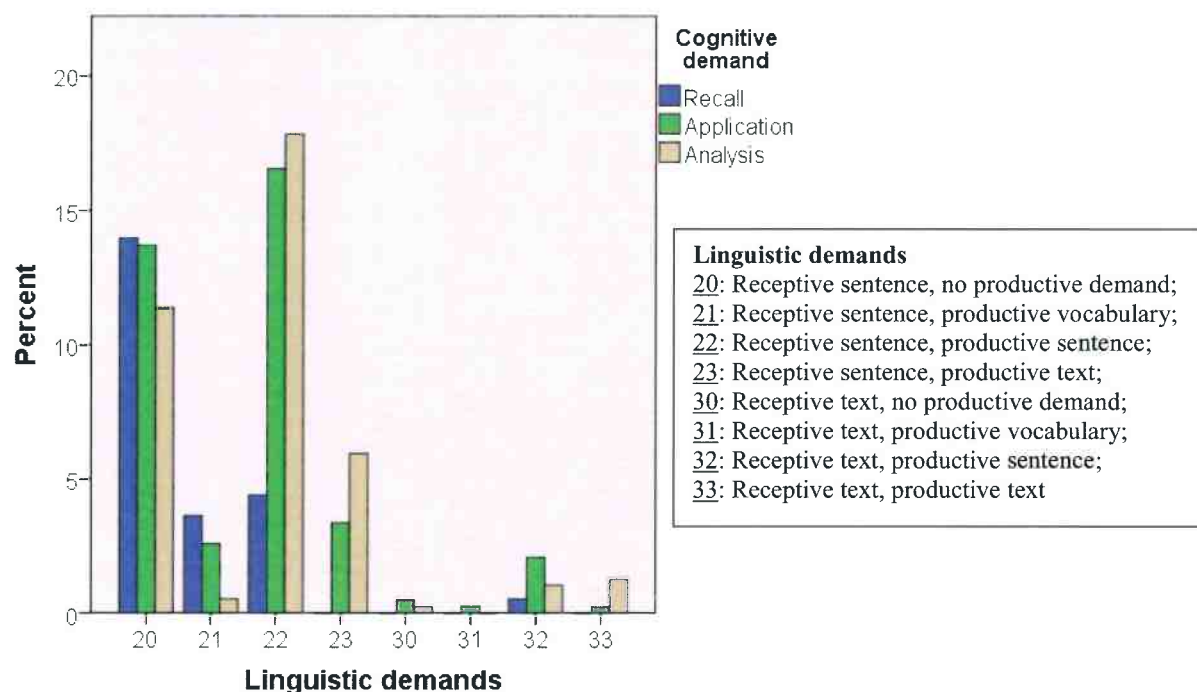


Figure 8. Bar chart showing distribution of Biology questions in HKDSE

For Biology, questions requiring understanding at the sentence level were predominantly represented (around 95% in total). Among these questions, there was quite an even split between those requiring no language production (39.02%) and sentence production (38.76%). However, while the former type of questions contained quite an equal proportions of different cognitive demands, the latter type of questions contained mostly application and analysis questions. In other words, there was a notable proportion of questions (around 35%) requiring application and analysis of knowledge on the sentence level both receptively and productively.

With Geography, as shown in Figure 9, all the questions contained a receptive sentence demand, meaning that understanding on the sentence level was also required in Geography. Similar to the Biology papers, there were quite a lot of questions necessitating no language production (46.61%), but these questions required more application than recall and analysis of knowledge. Additionally, compared to Biology, there were not as many questions requiring sentence production (14.97%); rather, the demand was shifted to the production of texts (30.79%). A majority of these text production questions involved more higher-order thinking and demanded some application and mostly analytical skills.

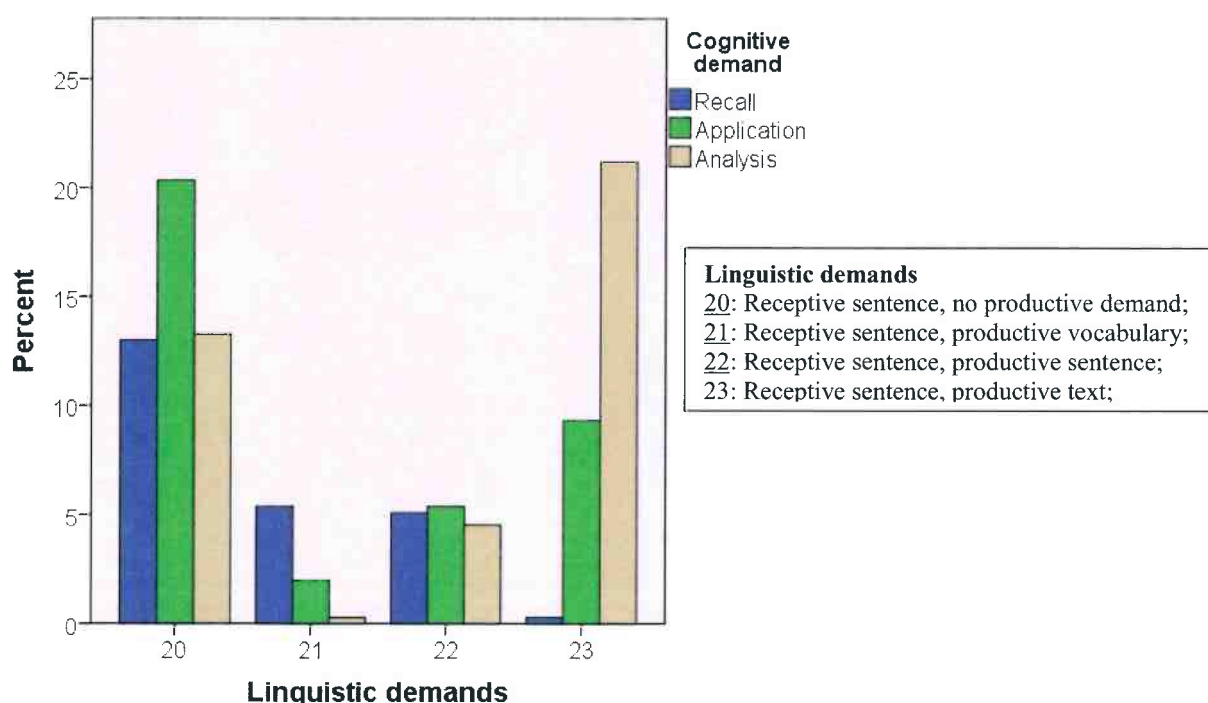


Figure 9. Bar chart showing distribution of Geography questions in HKDSE

#### *Summary of findings of Stage 1:*

This stage aimed to examine the current assessment practices of EMI education, particularly focusing on the cognitive and linguistic demands imposed on students. It also sought to compare the demands involved in different types of assessment and across different key

learning stages. The key observations are listed below:

- Regardless of grade levels and subjects, students are encountering both cognitive and linguistic demands in EMI assessments. Even at junior secondary levels, most questions require receptive language skills (i.e. reading and understanding the questions presented in sentences). Such linguistic demands appear to increase with grade levels.
- Regardless of subjects, there appears to be a progression in both cognitive and linguistic demands when students proceed to senior secondary levels. In particular, most questions at junior secondary ask for “recall” and “application” skills and require little language production, but those at senior secondary ask for “application” and “analytical” skills and require students to write sentences or short texts.
- Comparing formative and summative assessments, it is observed that the cognitive and linguistic demands imposed by questions in textbooks/workbooks and those in school-based examinations are rather similar. However, the questions in the public examination, HKDSE, appear to be more cognitively and linguistically challenging than those in senior secondary textbooks/workbooks.
- Comparing Biology and Geography, especially at senior levels, the questions in Geography impose higher productive linguistic demands and require more production at the text level than those in Biology.

### ***Results of Stage 2: Alignment among objectives, instruction and assessment***

The aim of this stage was to examine how objectives, instruction and assessment practices may interact or affect each other in particular EMI/CLIL school contexts. We will start this section by presenting our general observations across cases, and then illustrating such observations with two cases.

Across the cases, we could identify two types of EMI content subject teachers – the first type, which constituted the majority, was more content-oriented. Their lesson objectives, instructional activities, and teaching and learning materials mainly concerned the content knowledge or concepts. Although these teachers demonstrated their language awareness during pre- or post-lesson observation chats and the semi-structured interview (e.g. they understood students' language barriers; they were aware of some difficult words), they did not incorporate much explicit language instruction in their lessons. They mainly taught the key words of the topic, but seldom went beyond that to sentence or text level. When it came to assessment practices, this group of teachers was also content-oriented. Their assessment questions did not impose heavy linguistic demands on students. That may also explain why they did not seem to give a lot of feedback on students' language errors.

On the other hand, we identified a couple of teachers who seemed to be both content and language sensitive or aware. This is the second type of teachers. In addition to content coverage, some of their lessons had quite clear language objectives (e.g. helping students to read newspaper articles and extract key information, teaching students how to describe a graph in a short paragraph). With these language objectives in mind, some of their lesson time and instructional activities were devoted to language instruction or scaffolding, during which students' attention was temporarily drawn to learning the academic language. Then, the teachers also expected students to demonstrate their language skills in the assessment tasks, which imposed higher productive language demands (e.g. doing oral presentations, writing short essays).

To illustrate the two types of teachers identified more clearly, we have chosen two cases to demonstrate (and compare) their lesson objectives, instruction and assessment practices. To begin with, we will present their brief profile.

### ***Two illustrative cases: Miss A & Miss B***

The two cases, Miss A and Miss B, were chosen because both of them taught Science and four of their junior level Science lessons were observed in this study (S.3 class taught by Miss A and S.2 class taught by Miss B). The lessons observed in Miss A's class focused on the topic "Application of Enzymes" and those in Miss B's class focused on "Common Acids and Alkalis". Both teachers were experienced teachers (with over 15 years of teaching experience) and their teaching qualifications were similar (i.e. subject trained with teacher qualification). Perhaps the major difference lay in their school context – Miss A was teaching in a top school where all subjects (except Chinese-related ones) were taught in English, whereas Miss B was teaching in an average school which adopted rather complicated medium of instruction policies, in the sense that at junior secondary levels, some classes were EMI classes and some classes only learned Science and Mathematics through English. Hence, it would be reasonable to assume that in general, the academic ability and English proficiency of students in Miss A's class were higher than those in Miss B's class.

#### *1. Objectives of the lessons*

The first component we examined was lesson objectives. Based on the pre-lesson observation chats and also the lesson transcripts, we could infer the objectives of the lessons observed. Both Miss A and Miss B articulated mainly "content" objectives, and Miss A usually made the lesson objectives clear to her students at the beginning of her lessons (i.e. what they were going to learn and do in each lesson). However, in lesson 3 observed, Miss A explicitly

highlighted a language-related objective, *'Now we need to learn two things today. The first thing is to describe the result from the graph ...'* This may be regarded as both content and language objectives, as describing the results from a graph involves students' analytical ability (e.g. interpreting the graph and identifying different stages and the relationship between variables) and linguistic skills, particularly certain typical sentence patterns (e.g. *'As the temperature increases/decreases, the rate of reaction increases/decreases/remains unchanged'*).

The lesson objectives set also echoed the teachers' perception of their students' capacity in coping with EMI education. For instance, despite teaching in a good school, Miss A still noticed some learner diversity and she said she would apply separate strategies for students with different levels. For less competent students, Miss A focused more on knowledge consolidation and lower-level language instruction (e.g. spellings, pronunciation), while for more competent students, she focused more on higher-level training including sentence structures of a scientific writing. Similarly, Miss B also noticed diversity among her students and she believed that content subject teachers like her might sometimes need to shift her lesson focus from content-teaching to English-improving. In the post-lesson interview, Miss B did share a few strategies that she would adopt, including teaching pronunciation of key words, explicit instructions on sentence patterns (e.g. compare and contrast), reading the textbook together with students, using concept maps to help students summarise the lesson, and repeating key concepts and phrases. However, due to time constraints (the observed lessons were scheduled towards the end of the semester), Miss B did not apply these strategies during classroom observations.



## 2. Instruction in the observed lessons

In our data analysis procedures, different episodes of lessons were first categorised as “regulative” or “instructional”, and for those “instructional” episodes, they were further divided into “content-oriented” or “language-oriented” episodes. For both teachers, the percentage of “regulative” register (calculated based on the total number of words in the lessons) ranged from 6.9% to around 30%, with the mean being 14% and 24% for Miss A and Miss B respectively. The rather high percentage of regulative register in both cases was mainly due to the experiments conducted in the lessons, in which teachers needed to give a lot of instructions and guidance to manage students’ behaviour.

Regarding “instructional” register, which is the main focus of this study, “content-oriented” episodes constituted an average of 71% and 95% (out of the total number of words in content-oriented episodes) for Miss A and Miss B respectively. This reveals that there were more “language-oriented” episodes in Miss A’s lessons. In particular, “language-oriented” episodes occupied 26% and 67% in lessons 1 and 3 of Miss A’s lessons respectively. The exceptionally high percentage of language teaching in lesson 3 actually corresponded to Miss A’s objectives for that lesson (i.e. to describe the result from the graph). These will be further described below.

When we analysed the content-oriented episodes in detail, according to the different levels of cognitive demands, we observed that there was some spread across the different cognitive levels in Miss A’s lessons, with a mean of 40% , 54% and 6.8% for *recall*, *application* and *analysis* levels respectively. On the other hand, Miss B’s lessons mainly focused on *recall* skills (87%), with some attention paid to *application* (13%) but none to *analysis*. Such differences regarding cognitive demands may be attributed to the different grade levels of the



students (S.3 in Miss A's class vs S.2 in Miss B's class), the different topics involved ("Application of enzymes" for Miss A vs "Common acids and alkalis" for Miss B), and the general academic ability level of the students in the two schools.

Regarding the different levels of language-oriented teaching (i.e. lexico-grammar, sentence, and text), it was observed that for both teachers, the majority of language teaching episodes focused on teaching vocabulary or grammar (70% for Miss A and 87% for Miss B). Very often, the teacher would provide short definition or brief explanation of grammar items for the students. Such word teaching strategies were highlighted by both teachers in the interviews, and were also commonly observed in other EMI/CLIL literature (e.g. Koopman et al., 2014).

As one key objective in Miss A's third lesson was 'to describe the results from a graph', some language-oriented episodes in Miss A's lessons focused on sentence patterns or even text writing (i.e. the paragraph describing the results of the experiment). This constituted a rather high percentage of sentence and text teaching in lessons 3 and 4.

Such explicit instruction of language to address a particular type of question could be attributed to Miss A's awareness of the important role played by language in assessment. To help students complete assessment questions, Miss A often adopted "worked-example strategy", which involved step-by-step illustration of the model answer to Science/Biology problems in class. The language-oriented instruction observed in lessons 3 and 4 was a good illustration of such a strategy.

### 3. *Assessment practices*

As discussed in the Methodology section, the teachers' assessment practices can be divided into three main types, namely oral formative assessments (typically represented by teachers' questions in the lessons), written formative assessments (questions in worksheets or textbooks completed as homework) and written summative assessments (formal quizzes, tests, examinations). These will be discussed in this section.

#### (i) Oral formative assessments

We analysed teachers' questions according to the different cognitive or linguistic demands, so as to examine whether they aligned with the lesson objectives and instruction. Comparing questions about content and those about language, there are more questions about content (out of the total number of questions asked, only 10% and 3% of the questions asked by Miss A and Miss B were on language). Such a trend is perhaps not surprising, if we consider the proportion of their content-oriented and language-oriented teaching episodes.

Regarding the cognitive aspect, we observed that the dominant type of questions asked by both teachers was *recall* (21% in Miss A's lessons and 48% in Miss B's lesson), though there were similar percentage of *application* questions in Miss A's lessons too (16%, compared with 5% in Miss B's lessons). Analysis questions were rarely asked, with only 5% in Miss A's lessons. Such a spread of questions corresponded to the instructional foci of the lessons observed.

However, the situation was slightly different when analysing questions about language. It could be recalled that a certain proportion of Miss A's teaching was devoted to language teaching (particularly at the sentence and text levels). Yet, an overwhelming majority of

questions she asked about language focused on the lexico-grammar level (usually asking the meaning or part of speech of a word). This is probably because it would be difficult for teachers to assess students' understanding at sentence or text level, which could be better demonstrated through writing.

## (ii) Written assessment

Miss A

We could only gather a formative assessment task from Miss A and her students. That task is a take-home written assignment, which included one graph drawing question (to illustrate the results of an experiment) and three discussion questions based on the experiment. All these discussion questions require *application* skills, and two of them asked students to read questions in sentences, and express their answers in sentences. The remaining one asked students to produce a piece of short text to explain the results of the experiment. It is also this question which largely summarised what Miss A did in the four observed lessons. Hence, from the sample scripts collected, we analysed students' answers and Miss A's marking practices in detail.

Miss A awarded 10 marks for this question, and all students sampled performed quite well, getting 8 to 10 marks. We would argue that such good results could be attributed to Miss A's explicit language teaching in lessons 3 and 4, which helped students to formulate their answers. When we examined students' scripts, we found that most of them had jotted notes next to the question, and some notes were related to how to structure their answers. For example, one student wrote '*Describe the graph (using data)*' and '*Describe (1) Explain (1) Describe (2) Explain (2) Describe (3) Explain (3)*'. When we read students' answers, we observed that most of them could produce the sentence patterns that Miss A talked about, e.g.

'From  $0^{\circ}$  to  $40^{\circ}\text{C}$ , as temperature increases, rate of reaction increases'. All these demonstrated the effectiveness of Miss A's instruction in the lessons.

However, when we examined Miss A's marking practices, we noticed that she seemed to focus more on the content, as she put ticks next to some key words/phrases in the scripts and then the total mark awarded corresponded to the number of ticks given. She would also give a general comment such as '*Very good*' and '*Your answer is very accurate*', but '*accurate*' here probably referred to accuracy of content. We did not find a lot of comments on students' language errors, probably because most students could produce rather well-formed and grammatical sentences. In the interview, Miss A also said that her grading rubrics were more content-oriented. For short questions, marks would not be deducted from language errors. Only if students hit all the key points but missed out on the language in long questions, one mark would be deducted. Such grading practices were confirmed by Miss A's students in the student interviews. Some students admitted that as language errors were less important in summative assessment, they tended to focus more on the content (e.g. keywords). This is perhaps a good illustration of the backwash effect of assessment on students' learning behaviour.

Miss B

From Miss B's class, we managed to collect both formative and summative assessment. For the former, it was a unit exercise in the workbook, which consisted of five True/False questions, five multiple choice questions and three structured questions. When analysing these questions in detail, 40% required *recall* skills, 50% asked for *application* skills and the remaining 10% required *analytical* skills. Regarding linguistic demands, 77% of the questions did not require any language production (e.g. True/False questions, multiple choice

questions and some parts of the structured questions). Around 10% of the questions asked students to produce vocabulary or sentences, and 13% asked students to write a short piece of text. Hence it seems that the linguistic demands of the written formative assessment were not particularly high.

In the final examination paper (i.e. summative assessment), one structured question was about the topic Miss B taught in the observed lessons. That question included 7 parts, totaling 8 marks (i.e. most parts were awarded one mark). Hence, it is not surprising to see that students were not expected to write a lot when attempting those questions (50% required vocabulary, 37.5% required sentences and the remaining did not involve any language production). In terms of cognitive demands, over 60% of the marks were related to recall skills and the remaining on application skills. Hence, from this particular structured question, it seems that the cognitive and linguistic demands imposed on students were not very high. The relatively low cognitive and linguist demands in assessments are probably due to the teachers' awareness of students' capacity, especially in relation to the potential language barrier. In the interview, Miss B mentioned that to help students understand the questions, sometimes graph would be used and long questions tended to be shorter. She admitted that these would allow to students to learn how to write their responses in English, but she was also aware that avoiding longer writing may prevent students from learning how to write more complete responses. However, in face of students' diversity, such a dilemma seems inevitable.

From the small number of samples collected, students' performance varied. In the unit exercise, as students were not required to produce much language, their different results were largely due to their understanding of the key concepts. For the only productive text-level

question (i.e. *‘Describe how a person can prepare a red cabbage extract’*), some students managed to write a rather coherent text, probably because they referred to the textbook, while some others could only write some incomplete sentences with misspelling of some key words (e.g. pestle).

Similarly, in the examination paper, students were not required to write much. Most of them simply wrote some key words or phrases to address the questions. For example, in response to the question *‘State two safety precautions that Peter should take when doing the experiment’*, most students simply wrote such verb phrases as *‘wear safety goggles’*.

In both the formative and summative assessment tasks, Miss B appeared to focus more on the content when marking students’ work. She acknowledged her content-oriented marking practices in the interview. Similar to Miss A, she usually put ticks next to the target key words or points and then awarded marks. When there were misspellings or incomplete content, Miss B would use symbols to indicate them (e.g. circle the misspelt words; put “...” after the answer). Written feedback was rarely seen in the collected sample work.

#### *Summary of the findings of Stage 2:*

As stated at the beginning of this section, through examining the 12 case teachers, Stage 2 of this study identified two groups of EMI/CLIL teachers – one group tends to be more content-oriented, and the other attempting to incorporate more language teaching or scaffolding into their lessons and also assessment practices. Such key findings are in line with the results of previous studies (e.g. Walker, 2001; Tan, 2011).

If we consider the alignment among objectives, instruction and assessment, we actually

observed strong alignment in both groups of teachers – those who were more content-oriented focused more on the content knowledge when setting their lesson objectives, designing their instructional activities, assessment practices and marking rubrics; while those who were more language-aware paid more attention in the three components. However, if we consider the “dual” goal of EMI/CLIL programmes and also the interplay between cognitive and linguistic demands observed in assessments (see the findings of Stage 1), we would argue that the second group of teachers may better prepare their students to cope with the challenges in CLIL assessment and hence to achieve the dual goal.

### ***Results of Stage 3: Design and try out of assessment tasks***

In this stage, the research team and content subject teachers designed an informal assessment paper with reference to the theoretical framework (Figure 1), so as to include questions with different levels of cognitive and linguistics demands. Analysis of students’ performance on those different types of questions would allow the diagnosis of their learning in cognitive and linguistic dimensions. Tables 1 and 2 show students’ performance in the IS and Geography test respectively. It should be noted that the percentages in the tables represent students’ general performance on different types of questions (e.g. in Table 1, students obtained 59.2% of the mark for recall questions requiring understanding of sentences but no language production in the IS test).

Table 1. Students' performance on questions with different demands (IS test)

| <b>Cognitive demand</b> | <b>Receptive linguistic demand</b> | <b>Productive linguistic demand</b> | <b>Mean percentage of marks obtained</b> |
|-------------------------|------------------------------------|-------------------------------------|--|
| Recall                  | Sentence                           | No production                       | 59.2%                                    |
|                         | Sentence                           | Vocabulary                          | 58.0%                                    |
| Application             | Sentence                           | No production                       | 86.7%                                    |
|                         | Text                               | Vocabulary                          | 52.5%                                    |
|                         | Text                               | Sentence                            | 60.1%                                    |
| Analysis                | Sentence                           | No production                       | 80.0%                                    |
|                         | Text                               | Sentence                            | 28.3%                                    |
|                         | Text                               | Text                                | 27.2%                                    |

Table 2. Students' performance on questions with different demands (Geography test)

| <b>Cognitive demand</b> | <b>Receptive linguistic demand</b> | <b>Productive linguistic demand</b> | <b>Mean percentage of marks obtained</b> |
|-------------------------|------------------------------------|-------------------------------------|--|
| Recall                  | Sentence                           | No production                       | 74.6%                                    |
|                         | Sentence                           | Vocabulary                          | 51.5%                                    |
|                         | Sentence                           | Sentence                            | 70.4%                                    |
|                         | Sentence                           | Text                                | 44.6%                                    |
| Application             | Sentence                           | Vocabulary                          | 39.1%                                    |
|                         | Sentence                           | Sentence                            | 14.6%                                    |

From the descriptive statistics shown in Tables 1 and 2 (together with the results of inferential statistical tests), the following trends were observed:

- (i) It appears that students were encountering some challenges in the cognitive aspect in the Geography test, but to a less extent in the Science one. When keeping the linguistic demands constant (e.g. Recall-Sentence-No production), students' performance in the Science test did not decline significantly. In fact, their performance on application and analysis questions (especially those without any



language production) was quite good (scoring over 80%). However, the picture was quite different in the Geography test. When keep the linguistic demands constant, students' performance declined with increasing level of cognitive demands (e.g. from over 50% for Recall-Sentence-Vocabulary to 39% for Application-Sentence-Vocabulary; from 70% for Recall-Sentence-Sentence to 15% to Application-Sentence-Sentence). However, such a finding is bit contradictory to what the students reported in the interviews – those who took the Science test said they encountered some difficulties in the test, especially when tackling the structured questions about “destarching”, whereas such an issue was not reported by the students taking the Geography test.

- (ii) When keeping the cognitive demands constant, students' performance declined with increasing linguistic demands. This applied to both content subjects.
  - In the Science test, when the cognitive demand is *recall*, raising the linguistic demand from no production to vocabulary production does not make a difference in students' performance. However, when it comes to the *application* and *analysis* levels, the results seem to suggest that raising receptive linguistic demand from sentence to text, coupled with increasing productive linguistic demand might pose hindrance to students' performance. For example, at the *analysis* level, students could score around 80% if they were asked to read sentences without any language production; but when they were asked to read a piece of short text and then expressed their answers in sentences or texts, their performance dropped significantly to less than 30%.
  - Similarly, in the Geography test, students' performance declined with increasing linguistic demands for the same level of cognitive demands. For example, at the

*recall* level, students' performance dropped significantly from around 75% at no production level to 50% and 45% for productive vocabulary and text level respectively; at the *application* level, students' average score dropped significantly from nearly 40% to around 15% when they were asked to produce sentences instead of vocabulary. However, students' performance on questions requiring the combination of Recall-Sentence-Sentence skills did not follow the general trend, as it was as good as that on Recall-Sentence-No production questions.

- In the stimulated recall interviews, some student interviewees did mention some challenges related to the linguistic aspect. One student mentioned that in the Geography test, he encountered some problems when reading the fill-in-the-blanks questions or statements. When this happened, he tried his best to think about some possible words that may fit into those blanks to complete the sentences. Another student mentioned that when attempting the essay type question, he tried to refer to the pictures given to guess what the question was about and how he could answer the question. One student who took the Science test commented on how he attempted the structured questions, especially those parts asking students to write about the results and conclusion. He said the Science teacher had taught them some typical phrases and sentence patterns when reporting the results and drawing conclusion. At the same time, he would also refer to the textbook and learned about the relevant sentence patterns.

In general, by designing and analysing students' performance on questions with different types of cognitive and linguistic demands in assessment tasks, teachers could better diagnose students' strengths and weaknesses in both cognitive and linguistic aspects. Such analyses

could then inform teachers of their instructional activities (e.g. whether they need to reinforce the teaching of some concepts, whether they need to incorporate more language scaffolding to help students read questions or to express their ideas in sentences or short texts).

## **(I) Conclusions and Recommendations**

This three-phase study aims to provide a more comprehensive understanding of assessment issues in EMI/ CLIL education, by analysing the cognitive and linguistic demands of assessment questions, investigating whether teachers' instructional practices align with the dual goal and demands of assessments, and exploring how to design valid assessment tasks in EMI education in Hong Kong.

Our analysis of over 8500 questions from different types of assessment and different grade levels confirms the integral role played by language in assessment of content knowledge, and such role appears to become more significant when students proceed to senior secondary level. Our results also reveal some noticeable gaps between formative and summative assessments, and between junior and senior secondary level, both in cognitive and (productive) language demands. Students are likely to face more cognitively demanding and linguistically challenging questions in senior secondary summative assessments, particularly the high-stakes public examination.

Such cognitive and linguistic demands in assessments then raise concerns about whether EMI/CLIL students are supported by their content subject teachers. Our multiple-case study shows that only some content subject teachers attempt to have both content and language objectives in their lessons and incorporate explicit language scaffolding to prepare students for the content and linguistic challenges in assessments.

With these major findings of the study, the following recommendations are made for policy makers, school administrators and teachers:

- ▶ It is interesting to notice that while assessment questions impose both cognitive and linguistic demands on students, the marking rubrics place emphasis on the cognitive aspect only. Therefore, the examination authority or school administrators may consider putting more emphasis on the language aspect (e.g. in the form of communication marks), so that both teachers and students are motivated to pay more attention to language.
- ▶ There have been more and more professional development programmes for EMI/CLIL content subject teachers. Most of them focus on teachers' academic language awareness and pedagogical practices. While these are definitely important, perhaps another direction of these programmes could be on teachers' assessment awareness and assessment practices, so as to raise EMI teachers' awareness of how they can assess their students' learning progress and difficulties in a valid way, especially considering the fact that students' language proficiency seems to mediate their expression of content knowledge.
- ▶ There are several recommendations for designing EMI/CLIL assessments:
  - The findings of our study have validated the usefulness of the theoretical framework for analysing assessment questions in terms of cognitive and linguistic demands (Figure 1). The essence of this framework is the attention paid to the demands in both dimensions, so that teachers can be more aware of the distribution of questions with different levels of demands in their assessment tasks.
  - The theoretical framework is also a useful tool for teachers to better diagnose students' learning progress and difficulties. In particular, teachers can better understand

whether students are encountering language barriers by examining students' performance on questions targeting at the same cognitive level but with varying linguistic demands. For example, in our analyses of existing questions from different sources, it seems that higher cognitive demands usually come with higher linguistic demands (e.g. students were asked to read and write sentences or even short texts to demonstrate their analytical skills). Such a design may make it difficult to diagnose students' learning difficulties. It may be good to include a few more multiple-choice questions targeting at analytical skills. As this type of question usually imposes lower linguistic demands on students, teachers can better estimate students' understanding of the concepts, before engaging students to take up both cognitive and linguistic challenges.

- In some cases, especially at senior secondary level, it may be inevitable that some questions have to be presented in the form of a short text. To alleviate the potential linguistic hindrance students encounter, teachers may need to make sure that the language use in the scenario is as clear and simple as possible. Sometimes, they may consider adding pictorial cues as support.
- Considering the linguistic demands imposed by EMI/CLIL assessments, content subject teachers may need to pay more attention to the language aspect in lessons. They are strongly encouraged to consider including some language objectives relevant to the content objectives in their lessons, and then incorporate more language scaffolding when delivering their lessons. One useful strategy to incorporate language scaffolding, as identified in this study, is that teachers work on some assessment questions together with the students in lessons, so as to serve as demonstration or modelling. For instance, when answering questions in workbooks or worksheets, teachers can invite students to try to formulate the answer to some questions,

especially those challenging ones. Teachers can then demonstrate how to paraphrase and organise students' ideas with proper academic language. This not only helps to achieve the “dual goal” of EMI/CLIL education, but also better prepares students to overcome the challenges in assessments.

## Bibliography

- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain*. New York: David McKay Co Inc.
- Cenoz, J., Genesee, F. & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, 35(3), 243-262.
- Choi, P. K. (2003). The best students will learn English: Ultra-utilitarianism and linguistic imperialism in education in post-1997 Hong Kong. *Journal of Education Policy*, 18(6), 673-694.
- Coyle, D., Hood., P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.
- Education Bureau. (2006). *Further evaluation on the implementation of the medium of instruction guidance for secondary schools: Final report (2002-2004)*. Hong Kong: Government Printer.
- Gablasova, D. (2014). Issues in the assessment of bilingually educated students: Expressing subject knowledge through L1 and L2. *The Language Learning Journal*, 42(2), 151-164.
- Heine, L. (2014). Models of the bilingual lexicon and their theoretical implications for CLIL. *The Language Learning Journal*, 42(2), 225-237.
- HKEAA (Hong Kong Examination and Assessment Authority). (2013). *2013 HKDSE Entry Statistics*. [Online] Retrieved December 23, 2014, from [http://www.hkeaa.edu.hk/DocLibrary/HKDSE/Exam\\_Report/Examination\\_Statistics/dseexamstat13\\_2.pdf](http://www.hkeaa.edu.hk/DocLibrary/HKDSE/Exam_Report/Examination_Statistics/dseexamstat13_2.pdf)
- Hönig, I. (2010). *Assessment in CLIL: Theoretical and Empirical Research*. Saarbrücken: VDM Verlag Dr. Müller.

- Hughes, A. (2003). *Testing for language teachers* (2<sup>nd</sup> ed.). Cambridge, U.K: Cambridge University Press.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Koopman, G. J., Skeet, J., & de Graaff, R. (2014). Exploring content teachers' knowledge of language pedagogy: A report on a small-scale research project in a Dutch CLIL context. *Language Learning Journal*, 42(2), 123-136.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212-218.
- Lazaruk, W. (2007). Linguistic, academic, and cognitive benefits of French immersion. *Canadian Modern Language Journal*, 63(5), 605-628.
- Lin, A. M. Y., & Wu, Y. (2015). 'May I speak Cantonese?' – Co-constructing a scientific proof in an EFL junior secondary science classroom. *International Journal of Bilingual Education and Bilingualism*, 18(3), 289-305.
- Lo, Y. Y. (2014a). Collaboration between L2 and content subject teachers in CBI: Contrasting beliefs and attitudes. *RELC Journal*, 45(2), 181-196.
- Lo, Y. Y. (2014b). L2 language learning opportunities in different academic subjects in content-based instruction – Evidence in favour of “conventional wisdom”. *Language and Education*, 28(2), 141-160.
- Lo, Y. Y., & Lin, A. M. Y. (2014). Designing assessment tasks with language awareness: Balancing cognitive and linguistic demands. *Assessment and Learning*, 3, 97-119.
- Lo, Y. Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, 84(1), 47-73.
- Marsh, H. W., Hau, K. T., & Kong, C. K. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared



- with Chinese) for Hong Kong students. *American Educational Research Journal*, 39(3), 727-763.
- Massler, U., Stotz, D., & Queisser, C. (2014). Assessment instruments for primary CLIL: The conceptualisation and evaluation of test tasks. *The Language Learning Journal*, 42(2), 137-150.
- Mehisto, P. (2008). CLIL counterweights: Recognising and decreasing disjuncture in CLIL. *International CLIL Research Journal*, 1(1), 93-119.
- Nikula, T., Dalton-Puffer, C., & Llinares, A. (2013). CLIL classroom discourse: Research from Europe. *Journal of Immersion and Content-Based Language Instruction*, 1(1), 70-100.
- Orlich, D. C., Harder, R. J., Callahan, R. C., Trevisan, M. S., & Brown, A. H. (2013). *Teaching strategies: A guide to effective instruction (10<sup>th</sup> ed.)*. Belmont, CA: Wadsworth Cengage Learning.
- Pérez-Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15(3), 315-341.
- Schleppegrell, M. (2004). *The language of schooling: A functional linguistics perspective*. New York: Routledge.
- Shaw, S. (2012). International assessment of Geography through the medium of English: Analysing the language skills required. In P. Charzyński, K. Donert & Z. Podgórski (eds.), *Bilingual teaching - globalization, regional Geography and English integration* (pp. 24-44). Toruń, Poland: Association of Polish Adult Educators.
- Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly*, 10(4), 452-475
- Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly*,

27(4), 627–656.

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.

Trent, J. (2010). Teacher identity construction across the curriculum: Promoting cross-curriculum collaboration in English-medium schools. *Asia Pacific Journal of Education*, 30(2), 167-183.

Tsui, A. B. M. (2004). Medium of instruction in Hong Kong: One country, two systems, whose language?" In J. Tollefson & A. B. M. Tsui (Eds.), *Medium of instruction policies: Which agenda? Whose agenda?* (pp. 97-106). N.J.: Lawrence Erlbaum.

Walker, E. (2011). How 'language-aware' are lesson studies in an East Asian high school context? *Language and Education*, 25(3), 187–202.

Appendix 1. Test papers designed and tried out in Stage 3

*\*Remark: These test papers may not be made access to the general public, since some questions were extracted or adapted from the workbook and question bank of textbook publishers.*

---

**S.2 INTEGRATED SCIENCE**  
**The necessary conditions for photosynthesis**

Time Allowed: 35 minutes

Full Marks: 30

---

**Answer ALL questions on the answer sheet provided.**

**I. Multiple-choice questions (5 marks, 1 mark each)**

**Choose the best answer and write its corresponding letter in the box.**

1. Which of the following turns iodine solution blue-black?

- A. Carbon dioxide.
- B. Butter.
- C. Potato.
- D. Water.

C

2. What happens to a green plant when we destarch it?

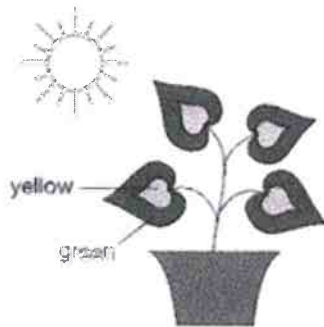
- (1) It carries photosynthesis.
- (2) It consumes stored starch.
- (3) It takes in carbon dioxide and releases oxygen.

- A. (2) only.
- B. (1) and (2) only.
- C. (1) and (3) only.
- D. (1), (2) and (3).

A

3. When we investigate whether light is needed for photosynthesis, which of the following set-ups is the most suitable?

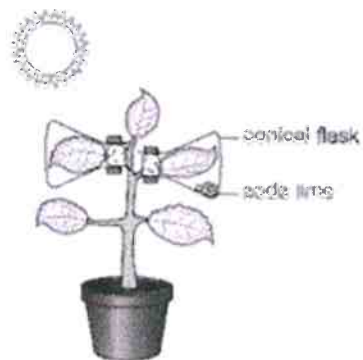
A. Destarched plant with variegated leaves under sunlight.



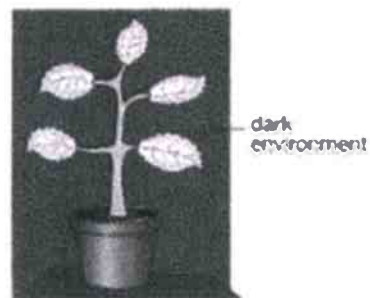
B. Destarched green plant under sunlight



C. Destarched green plant under sunlight

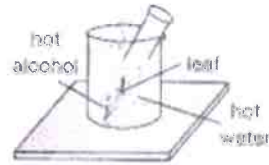


D. Destarched green plant in the dark.



B

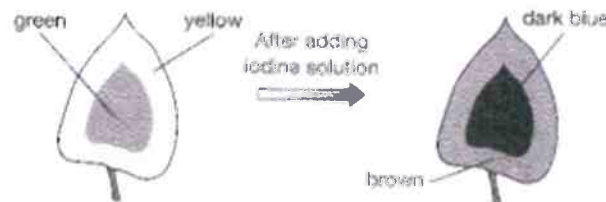
4. In an experiment of testing for the presence of starch in a leaf, what is the purpose of putting the leaf in a test tube with hot alcohol?



- A. Killing the leaf cells.
- B. Removing starch.
- C. Removing chlorophyll.
- D. Dissolving the protective layer on the leaf surface.

C

5. The diagram below shows an experiment using a variegated leaf to study a condition necessary for photosynthesis. What can we conclude from the result?



- A. Chlorophyll is necessary for photosynthesis.
- B. Water is necessary for photosynthesis.
- C. Sunlight is necessary for photosynthesis.
- D. Carbon dioxide is necessary for photosynthesis.

A

## II. Fill in the blanks (5 marks, 1 mark each)

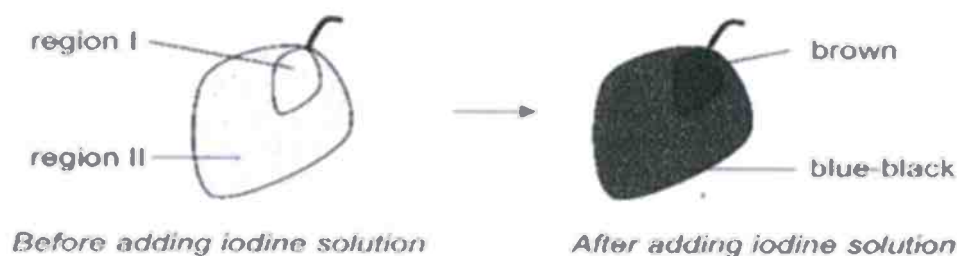
Fill in the blanks with the most suitable words.

1. Green plants can use carbon dioxide and water to make their own food.
2. Green plants carry out photosynthesis to produce starch and oxygen.
3. Plants cannot carry out photosynthesis in the dark.
4. Variegated leaves refer to the leaves that have green and non-green parts.
5. Iodine solution can be used to test for the presence of starch in food.

### III. Structured Questions (20 marks)

Answer the following questions. Bonus marks will be given to accurate and appropriate language use in the answers.

1. (8 marks) A plant with variegated leaves was left in the dark for 48 hours. The plant was put under bright light for 4 hours, and then a leaf was removed from the plant. Region I and II of the leaf are yellow and green respectively. Several hours later, it was put into boiling water for two minutes and then soaked in hot alcohol for 10 minutes. Finally, the leaf was washed with hot water. A few drops of iodine solution were added to the leaf as shown below. Region I was brown in colour and region II was blue-black in colour.



- (a) Why was the plant put in the dark for 48 hours? (1 mark)

Because the plant needed to be destarched

---

- (b) What was the purpose of soaking the leaf in hot alcohol? (1 mark)

The purpose of soaking the leaf in hot alcohol was to remove the chlorophyll from the leaf.

---

- (c) What was the purpose of washing the leaf with hot water after it was soaked in hot alcohol? (1 mark)

The purpose of washing the leaf with hot water was to wash away the alcohol and soften the leaf.

---

- (d) What conclusion can you draw from the experiment? Explain briefly. (5 marks)
-

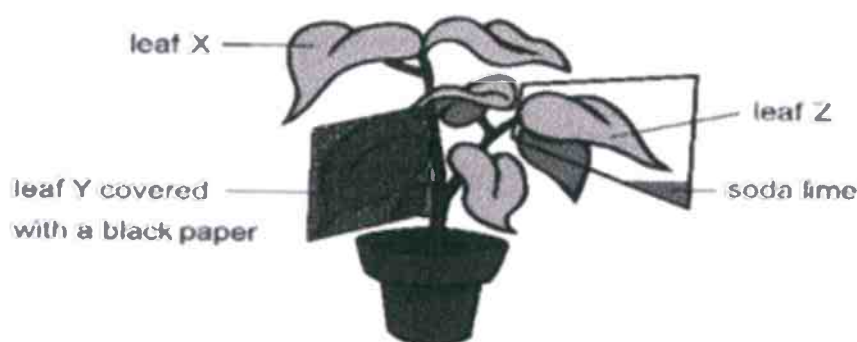
The results of the experiment indicate that region II contains starch(1) while region I does not(1).

This shows that region II which contained chlorophyll has carried out photosynthesis(1) while region I which did not contain chlorophyll has not carried out photosynthesis(1).

Hence, it can be concluded that chlorophyll is necessary (is needed) for photosynthesis(1).

---

2. (12 marks) The following experiment studies the condition necessary for photosynthesis. The green plant below has been put in the dark for two days before the experiment. Leaf Y is covered with a piece of black paper and leaf Z is contained in a sealed flask with a small amount of soda lime. After leaving the plant under the sun for a few hours, leaves X, Y and Z are removed for starch test.



(a) Leaves X, Y and Z are heated in boiling water before they are tested for starch. What is the purpose of doing so? (1 mark)

The purpose of heating the leaves in boiling water is to destroy the cell membranes of the leaf cells.

---

(b) What else should be done to the leaves before they are tested for starch? Why? (2 marks)

The leaves should be soaked in hot alcohol and then washed in hot water before they are tested for starch (1). Because the chlorophyll in the leaves needs to be removed. (1)

---

(c) Suggest one chemical which can be used to test for starch. (1 mark)

Iodine solution can be used to test for starch.

---

---

(d) What is the positive result of the test for starch? (1 mark)

The iodine solution turns/changes to/becomes blue-black.

---

(e) Write down the result of this experiment. (2 marks)

The iodine solution turns blue-black in leaf X (1) while it remains brown in leaves Y and Z (1).

---

(f) What conclusion can you draw from the experiment? Explain briefly. (5 marks)

The results of the experiment indicate that starch is present in leaf X (1) while it is absent in leaves Y and Z (1).

This shows that only leaf X has carried out photosynthesis (1) while leaves Y and Z have not (1).

Therefore, we can conclude that both light and carbon dioxide are necessary for photosynthesis (1).

---



## S.2 GEOGRAPHY

### What harmful effects do scientific farming methods bring?

Time Allowed: 25 minutes

Full Marks: 25

---

**Answer ALL questions on the answer sheet provided.**

#### **I. Multiple-choice questions (5 marks, 1 mark each)**

**Choose the best answer and write its corresponding letter in the box.**

1. Which of the following are the examples of scientific farming methods?

- |   |   |
|---|---|
| (1) Using chemical fertilizers and pesticides | (3) Applying GM technology                            |
| (2) Using drip irrigation                     | (4) Using traditional and simple tools to grow crops. |

- A. 1 and 4 only
- B. 2 and 3 only
- C. 1, 2 and 3 only
- D. 1, 2, 3 and 4

C

2. What are the negative effects of using too many pesticides?

- (1) Good insects are killed.
- (2) Land and water will be polluted.
- (3) Crops are contaminated.
- (4) Pests will become more resistant to pesticides.

- A. 1 and 2 only
- B. 1, 3 and 4 only
- C. 2, 3 and 4 only
- D. 1, 2, 3 and 4

D

3. Which of the following are the negative impacts of scientific farming methods?

- (1) Environmental pollution
  - (2) Soil degradation
  - (3) Soil erosion
  - (4) Disturbance of the natural ecosystem
- A. (1) and (3) only
  - B. (2) and (3) only
  - C. (1), (2) and (4) only
  - D. (1), (2), (3) and (4)

D

4. Why do some people oppose the growing of GM crops?

- (1) They may not be safe to consume.
  - (2) They may affect the natural environment.
  - (3) They need longer time to grow.
  - (4) The development of GM crops is against the laws of nature.
- A. 1 and 3 only
  - B. 2 and 4 only
  - C. 1, 2 and 4 only
  - D. 2, 3 and 4 only

C

5. What were the impacts of blue-green algae bloom in Tai Hu in 2007?

- (1) Water pollution
- (2) Dead of fish
- (3) Affecting navigation
- (4) Decreasing freshwater supply



The blue-green algae bloom in Tai Hu in 2007

- A. (1) and (4) only
- B. (2) and (3) only
- C. (1), (2) and (4) only
- D. (1), (2), (3) and (4)



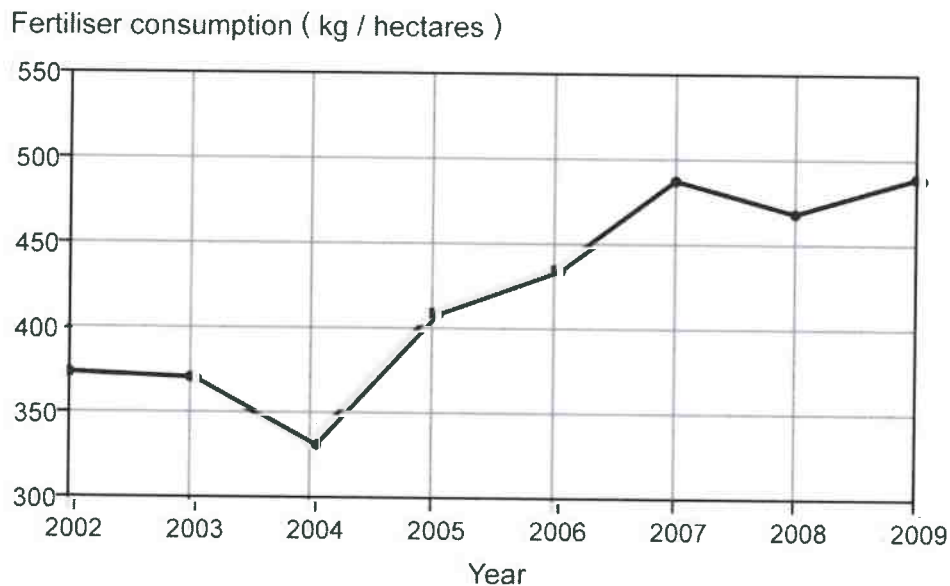
## II. Fill in the blanks (5 marks, 1 mark each)

Fill in the blanks with the most suitable words.

1. When excessive fertilizers are washed into rivers, plants and algae will grow rapidly and oxygen in the water will be used up.
2. Scientific farming methods increase farming areas and the productivity of each farm unit.
3. Farmers can use pesticides to control and kill pests.
4. The use of scientific farming methods requires huge investments in infrastructure, such as the building of dams and irrigation pipelines.
5. Rearing too many animals in semi-arid areas will lead to soil erosion and soil degradation.

### III. Structured Questions (7 marks)

Answer the following questions. Bonus marks will be given to accurate and appropriate language use in the answers.



**Fig. 1 Chemical fertilizer consumption in China**  
**2012**

Source: World bank,

- (a). Describe the trend of fertilizer consumption in China between 2002 and 2009. (2 marks)

The chemical fertilizer consumption in China had an increasing trend between 2002 and 2009. / It increased from 375 kg per hectares in 2002 to around 500 kg per hectares in 2009.

(Any 1)

- (b). Explain the trend. (2 marks)

Chemical fertilizers can *improve* the fertility of soil and *raise* farm productivity. As a result, the income of the farmers *increases* and they can *consume more* fertilizers. (Or other reasonable answers)

- (c). What are the negative impacts of excessive use of chemical fertilizers? (3 marks)

Too many fertilizers may *undermine* the productivity of farmland / *pollute* the soil, water and rivers / *destroy* soil structure / *affect* ecological balance / *threaten* the health of humans and wildlife (Any 3)

#### IV. Short essays (8 marks)

1. Refer to Figure 1. What are the negative effects caused by these improper uses of scientific farming methods?

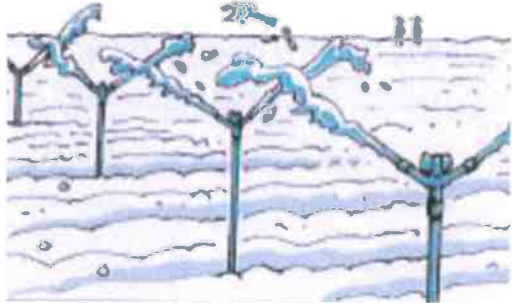


|  |  |
|--|--|
|   |  |
| Improper irrigation  | Open up of farmland in marginal land   |
|  |  |
| Overuse of fertilizers and pesticides  |  |

Figure 1

*As shown in figure 1, the negative effects caused by improper uses of scientific farming methods include the following:*

Firstly, improper irrigation may **lead to** the building up of salts in the soil and lower farm production.

Secondly, opening up more farmland in marginal land may **result in** soil erosion and soil degradation.

Thirdly, overuse of pesticides and fertilizers may **cause** water pollution. Living things in rivers may be killed. Pesticides may pollute crops. People may be poisoned when they eat the polluted crops.

*To sum up, although scientific farming methods may help improve productivity and increase the output of farming, they should be used properly to avoid the negative impacts.*