Assessment in CLIL: Test Development at Content and Language for Teaching Natural Science in English as a Foreign Language

Johanna P. LEAL*

Abstract

On-going bilingual programs without regard to needs analysis; little research on the actual effects of CLIL in Colombia and vague awareness or knowledge about the necessary considerations for effective CLIL programs, underpin the need to address a particular issue of curriculum as it is summative assessment. This small scale study takes place in a Natural Science class using a CLIL approach with third-grade students at A2 proficiency level who have been progressively immersed in a bilingual program at a private school in Bogotá, Colombia. Regularly scheduled tests were analyzed in order to identify suitable assessment items that simultaneously report on the content and language achievement in order to provide guidelines for test development that are aligned with the teaching goals, consistently measure students' progress, and facilitate teaching practices. This study entails a systematic examination of test items using formal item analysis to depict test validity from an assessment grid that integrates content, at different knowledge levels, CALP functions and cognitive skills. The study concludes that the assessment grid is a helpful tool to discriminate language and content achievement in the results of multiple-choice CLIL tests, by increasing teachers' understanding of the language demands of test items and the level of difficulty of content tasks.

Keywords: Science; young learners; CLIL; summative assessment; assessment frameworks; reliability.

* orcid.org/0000-0002-4070-1384. Universidad de La Sabana, Colombia. johannaleva@unisabana.edu.co

Received: 2016-09-04 / Sent for peer review: 2016-09-28 / Accepted by peers: 2016-11-08 / Approved: 2016-11-16

To reference this article in APA style / Para citar este artículo en APA / Para citar este artigo

Leal, J.P. (2016). Assessment in CLIL: Test development at content and language for teaching natural science in English as a foreign language. *Latin American Journal of Content and Language Integrated Learning*, 9(2), 293-317. doi:10.5294/laclil.2016.9.2.3

La evaluación en AICLE: el diseño de pruebas de contenido y lengua para enseñar ciencias naturales a través del inglés como lengua extranjera

Resumen

Los programas bilingües actuales carentes en cuanto a análisis de necesidades, la investigación insuficiente relacionada con los efectos de AICLE en Colombia, así como la poca conciencia y conocimiento acerca de las consideraciones necesarias de los efectos de AICLE, señalan la necesidad de enfocarse en un aspecto curricular particular como es el de la evaluación sumativa. El presente estudio a pequeña escala se realizó en una clase de ciencias naturales en la que AICLE es el enfoque seleccionado para la enseñanza a estudiantes de tercer grado con un nivel de competencia A2 y quienes se encuentran en un programa de bilingüismo progresivo en un colegio privado en Bogotá, Colombia. Se analizaron pruebas ordinarias para identificar preguntas de evaluación apropiadas que permitan reportar simultáneamente los logros en contenido y lengua, con el fin de construir lineamientos para el diseño de pruebas que estén alineadas con las metas de enseñanza, que midan consistentemente el progreso de los estudiantes y faciliten las prácticas de enseñanza. Este estudio implicó el análisis sistemático de las preguntas de las pruebas usando un análisis formal de preguntas para determinar la validez de las pruebas a partir de la aplicación de una matriz de evaluación que integra el contenido en diferentes niveles del conocimiento, el dominio cognitivo del lenguaje académico (DCLA) y las habilidades cognitivas. El estudio concluyó que la malla de evaluación es un instrumento útil para discriminar los logros en el aprendizaje de contenido y lengua en los resultados de pruebas de selección múltiple de AICLE, al facilitar e incrementar la comprensión de los profesores en relación con las exigencias de la lengua en las preguntas de las pruebas y el nivel de dificultad en cuanto a contenido.

Palabras clave: ciencias naturales; niños; AICLE; evaluación sumativa; marcos de evaluación; confiabilidad.

A avaliação na AICLE/CLIL: o desenho de provas de conteúdo e língua para ensinar ciências naturais por meio do inglês como língua estrangeira

Resumo

Os programas bilíngues atuais carentes, quanto à análise de necessidades, a pesquisa insuficiente relacionada com os efeitos da AICLE/CLIL na Colômbia bem como a pouca consciência e conhecimento sobre as considerações necessárias dos efeitos da AICLE/CLIL indicam a necessidade de se enfocar num aspecto curricular particular, como é o da avaliação sumativa. Este estudo, em pequena escala, foi realizado numa aula de ciências naturais na qual a AICLE/CLIL é a abordagem selecionada para o ensino a estudantes de terceiro grau com um nível de competência A2 e que se encontram num programa de bilinguismo progressivo num colégio particular em Bogotá (Colômbia). Analisaram-se provas ordinárias para identificar perguntas de avaliação apropriadas que permitam relatar simultaneamente as realizações em conteúdo e língua, a fim de construir lineamentos para o desenho de provas que estejam alinhadas com as metas de ensino, que meçam conscientemente o progresso dos estudantes e facilitem as práticas de ensino. Este estudo implicou a análise sistemática das perguntas das provas usando uma análise formal de perguntas para determinar a validade das provas a partir da aplicação de uma matriz de avaliação que integra o conteúdo em diferentes níveis do conhecimento, o domínio cognitivo da linguagem acadêmica (DCLA) e as habilidades cognitivas. O estudo concluiu que a grade de avaliação é um instrumento útil para discriminar os progressos na aprendizagem de conteúdo e língua nos resultados de provas de múltipla escolha da AICLE/CLIL, ao facilitar e aumentar a compreensão dos professores quanto às exigências da língua nas perguntas das provas e o nível de dificuldade com relação ao conteúdo.

Palavras-chave: AICLE/CLIL; avaliação sumativa; ciências naturais; confiabilidade; crianças; referenciais de avaliação.

INTRODUCTION

Colombia has fostered bilingualism through different projects and national policies: The General Education Law (Congreso de Colombia, 1994) , the Colombian Bilingual Project 2004-2019 (Ministerio de Educación Nacional, 2004), the Guide to National Standards for the Development of Foreign Language Competencies – Guía # 22 (Ministerio de Educación Nacional, 2006) the Law of Bilingualism (Congreso de Colombia, 2013) , and the English National Program 2015-2025 (Ministerio de Educación Nacional, 2013). Consequently, the implementation of bilingual programs has been developed by many private and public schools, developing an increasing interest for Content Language Integrated Learning – CLIL as a bilingual approach (McDougald, 2009).

This national tendency has led private and public schools to the implementation of programs without regard to learners' needs analysis, context characteristics, and required resources (Lugo-Vásquez, Fandiño-Parra, & Bermúdez-Jiménez, 2012). Additionally, little research has been conducted in Colombia regarding the actual effects of CLIL: one study was found at the university level (Otálora, 2009), one at elementary school level (Mariño, 2014), two regarding teachers' perceptions and experiences (Curtis, 2012a) (Curtis, 2012b) (McDougald, 2015) and two more related to the state of CLIL in Colombia (McDougald, 2009), (Rodríguez, 2011).

Hence, there is an urgency to initially focus on specific aspects of the curriculum that can provide information about the effectiveness of the program in the short term. Assessment is an alternative used to gather information about the teaching and learning process (Bailey, 1998) as well as a practice that is regularly part of school systems. It could open a door to initiate further studies that can lead to the comprehension of the results in both content and language competencies. Particularly, summative tests can become a useful tool if it is consciously conceived to measure what it is intended to measure, providing consistent results, and being practical enough to be assumed by content teachers under regular working conditions.

In this regard, this study developed an assessment grid adapted from two tools: the CLIL Matrix suggested by Coyle, Hood, & Marsh (2010) and

a conceptual framework proposed by the project Assessment and Evaluation in CLIL – AECLIL. The former tool sets the route of difficulty among content and language, reports on literature (Short, 1993; Coyle, Hood, & Marsh, 2010; Lo & Lin, 2014) and reveals how information provided by this Matrix support informed decisions by teachers (Coyle, Hood, & Marsh, 2010, p. 68). The latter test provides the theoretical assumptions to define and relate content, cognition and language skills. The assessment grid seeks to facilitate the process of sorting test items through a route that integrates cognitive and linguistic demands.

This study focused on determining to what extent this assessment grid of content and language demands provides a guideline for test development that aligns with the teaching goals, consistently measures students' achievement, and could be implemented under regular teaching conditions. This study entails a systematic examination of test items using Wesche's framework (1983 as cited in Bailey, 1998, p. 13) as the categories to classify items in the assessment grid and ensure test validity.

Finally, this small scale study aims at impacting curriculum development in approximately 175 bilingual schools officially registered in Colombia (Ministerio de Educación Nacional, 2009) by providing a guideline to design multiple-choice tests that simultaneously provides information about content and foreign language development. Valid and reliable assessment items can initially support content teachers in their process of lesson planning and material design as they are better informed about the content and language needs of their students.

METHOD

This study examined three tests that went through the research design. Firstly, a systematic design of tests using Wesche's framework (1983 as cited in Bailey, 1998, p.13) to place each test item in the assessment grid. Tests were collaboratively developed by a Content and Language Integrated Learning (CLIL) teacher and an English as a Foreign Language (EFL) teacher in order to ensure construct validity in terms of content and language. Secondly, an item analysis was carried out to determine the reliability of each item. Consequently, a report was built to elucidate the

items' validity and reliability and define the overall results of the test in terms of content and/or language achievement.

The framework provided by Mari Wesche (1983 as cited in Bailey, 1998, p.13) is a simple yet useful tool for examining tests in four parts: stimulus material, the task posed to the learner, the learner's response, and the scoring criteria. Particularly, this study focused on two aspects of Wesche's framework (1983 as cited in Bailey, 1998, p.13): the stimulus material to analyze test input in terms of language demands and the task posed to the learner to identify the content demands of each test item. The data provided by this framework allowed for the placement of test items in the assessment grid.

Assessment Grid

The main goal of this study comes from the concern that CLIL, as a dual-focus approach, requires assessment of students' achievement in both content and language components, so teachers can identify which area is interfering in students' learning. In order to reach this goal, this study has combined two theoretically-accepted tools: The CLIL Matrix suggested by Coyle, Hood, and Marsh (2010) and a conceptual framework proposed by the Evaluation and Assessment in CLIL Project (Quartapelle, 2012). The product of this integration is illustrated in Table 1.

Table 1. Assessment grid

<table>
<tr><td rowspan="2">Content Demands - Knowledge structure</td><td>High</td><td>Principles/relationships
Relationship between concepts – principles-processes - routines</td><td>Quadrant I
Defining
Identifying
Classifying
Describing...</td><td>Quadrant II
Applying
Explaining
Comparing
Analyzing...</td></tr>
<tr><td>Low</td><td>Concepts/classification
What? – Where? – Who? - When?</td><td>Quadrant III
Defining
Identifying
Classifying
Describing...</td><td>Quadrant IV
Applying
Explaining
Comparing
Analyzing...</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>Lower-order Thinking skills / CALP functions</td><td>Higher-order Thinking skills / CALP functions</td></tr>
<tr><td colspan="2">Language Demands</td></tr>
</table>

As seen in Table 1, it is clear that the CLIL Matrix provides the parameters to place the conceptual framework in four quadrants that make visible the interconnectedness among content and language demands. Each quadrant frames a particular connection of knowledge, thinking skills and the language necessary for its understanding. Accordingly, Quadrant I – QI – denotes all items that require high content demand at a low language level. Quadrant II– QII– describes items at the highest levels of content and language demands. In contrast, Quadrant III – QIII– corresponds to the lowest content and language demands. Finally, Quadrant IV – QIV – challenges students with high language levels to answer low content demanding questions.

In pedagogical terms, Coyle, Hood, and Marsh (2010) highlight that whilst QIII might build initial confidence in students, in CLIL is likely to be a transitory step on the way towards QII. However, the transition from QIII to QII or IV focuses on progression of individuals and the realization of their potential over time (2010, p. 44).

Context of the Study

This study took place at a private school that has established a bilingual program within the characteristics of an early partial immersion (Baker, 2006 as cited in Pacific Policy Research Center, 2010) in which students from age 5 or 6 have 50% of the curriculum taught through English as a Foreign Language – EFL during their elementary education. The program is at a stage of on-going implementation in which students currently in third grade have increased the number of subjects instructed in English since 2014 to date (2016) when they finally have 50% of their curriculum in English. This study focused on the evaluation of CLIL in science as it is the only content subject that is assessed by the national standard tests, has a relevant number of hours in the curriculum, and is the second most popular content subject taught in Colombian Bilingual Schools (McDougald, 2015).

In accordance with this context, bilingual teachers are mainly content specialists who have an upper-intermediate mastery of EFL. They have a tendency to be concerned more with the development of content competencies, ignoring language constraints that regularly affect

mixed-ability language learners in CLIL settings. Furthermore, administrators at this private school did not carry out a needs analysis to set specific guidelines for the implementation of CLIL as it is suggested by many authors (Coyle, Hood, & Marsh, 2010) (Butler, 2005, as cited in McDougald, 2015). The teachers themselves seem to have only vague awareness or knowledge about the considerations necessary to establish effective CLIL programs (Butler, 2005, as cited in McDougald, 2015).

Validation

Validation of the study was underpinned by the use of different sources of analysis in each phase. In phase I, the collaborative work done by the CLIL teacher and the EFL teacher, through individual and pair analysis systematically using Wesche's framework (1983 as cited in Bailey, 1998, p.13) and the assessment grid allowed certain degree of quality, that could be latter assessed during phase II.

In phase II item analysis was performed from three different perspectives that are commonly used to examine the quality of multiple-choice test on classrooms: Item Facility, item discriminability, and distractor analysis. The individual results and its analysis as a whole provided a holistic picture of each test item and determined whether those items were acceptable or not for the purpose of the study.

Item Facility is an index that represents the portion of students who answered each item correctly. It provides a source of analysis to help establish the level of difficulty claimed for each test item according to the assessment grid. In order to uncover the variability in skills and/or knowledge that is assumed to exist in a group of test-takers, a comparison of the good students and the poor students, in terms of how they perform each item, provides useful information in the discrete-point, norm-reference approach. Item Discrimination – I.D. examines test items in a more accurate way as it shows how the top scorers and the lower scorers performed on each item. These statistics allow you to determine whether the item with a low I.F. is actually difficult, or if other factors might influence the low rate of correct responses for that item. Point-Biserial correlation coefficient is the most appropriate tool suggested by Bailey to determine item discrimin-

ability. Finally, Distractor Analysis is a procedure specifically related to the multiple-choice formats. It shows how each individual distractor is functioning. An important aspect affecting the difficulty of multiple-choice test items is the quality of distractors. Some distractors, in fact, might not be distracting at all, and therefore serve no purpose. This approach assumes that there is some variability (Bailey, 1998, p. 134).

RESULTS

Three tests were analyzed in order to identify their characteristics in terms of language and content demands, and placed their items in the assessment grid with the intention to discriminate which of the two constructs required more instruction, or have been mastered by students.

By and large, tests items were mainly placed in QI and QII, suggesting that there is a high emphasis in assessment of content knowledge with low demand on language. Only Test Three had a valid item in QIV. This brings attention to the difficulty that may entail the design of low content demand questions with high language demands. The number of items that need revision varied from 1 to 3. A positive improvement was observed in the number of distractors that needed replacement. The assessment yielded useful categorization of items, in particular when they were related to each other in terms of content components.

Test One

This test was a diagnostic that contributed to the starting point as to how tests were initially developed. At the beginning of the school year, 89 third-grade students in five different classrooms took a 12-item multiple-choice test that had as a purpose to determine students' entry levels of content competencies according to the exit outcomes planned for second grade, and the corresponding foreign language understanding. This is shown in Table 2.

The CLIL teacher and the EFL teacher collaboratively wrote the questions; meanwhile they classified each test item in the assessment grid. The process of sorting out each item was supported by the Wesche's framework (see Appendix A) and resulted in the information showed in table 3.

It is evident from the assessment grid that the test focused on low language demands as items are mainly placed in QI and QIII, which could be explained due to the diagnostic intention of testing students who just started their school year and faced for the very first time this content class in a foreign language.

Table 2. Content and language components, Test One

Topic	Living things
Components	1. To describe, compare, and contrast living things and nonliving things. 2. To identify what living things need. 3. To classify living things according to the kingdoms. Based on the national standards released by the Colombian Ministry of Education and the school curriculum.
Language functions	Describing – Comparing – Contrasting – Classifying - Observing
Language Structures	It grows... It can move... It doesn't need food And – but
Vocabulary	Living things – Biotic factors Nonliving things – Abiotic factors: sand, rocks, water, sunlight, temperature, air. Life processes: growth – nutrition – respiration – sensation – excretion - reproduction Kingdoms: insects, mammals, reptiles, birds, amphibians, fungi, protists

Table 3. Assessment grid, Test One

<table>
<tr><td rowspan="2">Content Demands - Knowledge structure</td><td>High</td><td>Principles/relationships
Relationship between concepts – principles-processes - routines</td><td>Quadrant I
Identifying
Items: 3 – 4 – 8 – 10</td><td>Quadrant II
Explaining Item: 11
Comparing Item: 12</td></tr>
<tr><td>Low</td><td>Concepts/classification
What? – Where? – Who? - When?</td><td>Quadrant III
Identifying
Items: 1 – 2 – 5 – 6 – 7 – 9</td><td>Quadrant IV</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>Lower-order Thinking skills / CALP functions</td><td>Higher-order Thinking skills / CALP functions</td></tr>
<tr><td colspan="2">Language Demands</td></tr>
</table>

Accordingly, test items that depicted cognitive academic vocabulary were placed in QI or QII because they demanded more content knowledge while their language features were mainly illustrated or contextualized. Items 11 and 12 required students to understand complex sentences as well as related cognitive academic vocabulary with the specific concepts and processes of the content subject.

Test One had a total of 12 items: 4 placed in QI, 2 in QII, and 6 in QIII. The content of the items was focused on three different components that affected the analysis of the reliability among items. Both Item Facility (I.F.) and Item Discrimination (I.D.) (See Appendix B) showed acceptable values for most of the items, although two items were found to need revision: Items 4 and 11. However, 17% of distractors (See Appendix C) corresponding to items 1, 4, 5, 6, 8 and 11 needed to be revised. Special attention should be paid to students when they are taking the exam because there was a meaningful number of items, whose performance was affected by no or wrong answers following the item instructions. It is important also to notice that the first test did not include any item in QIV due to the teaching tendency to focus more on the content demands rather than the language demands.

Table 4 consolidates overall results around item 12. This review does not include items 4 and 11 because they were found to affect the overall performance. This table shows that 45% of students achieved the high content and language demands of QII. In this regard, only 30% of students answered correctly low content and language demands in QIII, and a similar percentage (29%) the high content at low language demands of QI. These findings show that items placed in the assessment grid do not depict the expected discrimination between content and language demands. This event might have been influenced by a few things, (a) the test is a diagnosis before instruction, (b) items measure different content components, and (c), items are not balanced within the assessment grid. Conclusions on these tests are twofold. First, test development needs to be enhanced by clarifying its purposes and content components. Second, students seem to need instruction in test-taking skills and academic language in order to understand test tasks.

Table 4. Results, Test One in the assessment grid

Q*	II	III						I		
Item	12	1	2	5	6	7	9	3	8	10
#**	40	29	38	18	22	26	26	29	28	20
%***	45	30						29		
Note:	*Quadrant in the assessment grid. **Number of students who answered correctly each item. Numbers in the other quadrants are taken from the set of students who answered correctly item in QII. ***n=89									

Test Two

Test Two was applied as an achievement measurement at the end of the first school term that lasted three months. In order to design Test Two, the CLIL teacher defined the content outcomes that were expected to be achieved and the EFL teacher identified the language components. Both are shown in Table 5.

Table 5. Content and language components, Test Two

Topic	Living things
Components	1. Make Assumptions based on observable evidence to answer questions. 2. Assume and test living thing's needs. 3. Identify common characteristics in living things. 4. Describe characteristics of living things, identify similarities, differences, and classify them according to them. Based on the national standards released by the Colombian Ministry of Education and the school curriculum.
Language functions	Describing – Identifying – Explaining – Classifying - Hypothesizing
Language Structures	Imperatives: Observe, choose, compare. Wh-Questions: Why, what Present Simple: are, is, do, have, belong Modal verbs: can, need
Vocabulary	Living things – Biotic factors: animals and their body parts. Nonliving things – Abiotic factors. Cell Types: Unicellular, Pluricellular, Multicellular, Eukaryotic, Prokaryotic. Domains: Protist, Fungi, Plantae and Animalia Kingdoms: insects, mammals, reptiles, birds, amphibians, fungi, protists

Table 6 shows that most of the items in the test included specific vocabulary of the subject such as types of cells, domains, and kingdoms. Items 2, 3 and 10 used basic interpersonal vocabulary. In regard to the difficulties yielded by the different content components assessed in Test One, Test Two involved a specific target component as it is identifying and classifying organisms in terms of domains and kingdoms. This is not the case of items 1 and 11, placed in QI because they demand an understanding of specific content terms such as scientific questions and hypotheses for general skills development of the content.

Table 6. Assessment grid, Test Two

Content Demands - Knowledge structure	**High**	**Principles/relationships** Relationship between concepts – principles-processes - routines	**Quadrant I** Identifying Items: 1 – 5 – 8 – 11 Classifying Items: 12	**Quadrant II** Explaining Item: 4
	Low	**Concepts/classification** What? – Where? – Who? - When?	**Quadrant III** Defining Items: 6 – 7 Identifying Items: 2 – 3 – 9	**Quadrant IV** comparing Item: 10
			Lower-order Thinking skills/ CALP functions	**Higher-order Thinking skills/ CALP functions**
			Language Demands	

Test Two analysis examined each of the 12 items in detail according to the assessment grid due to its emphasis on a specific content component. Five items placed in QI, 1 in QII, five in QIII, and one in QIV and described a test with better distribution of items compare to Test One which had more items in QIII and none in QIV. Additionally, it is worth noting, that items in Test Two had more specific content vocabulary, although its items had more context clues. Only item 10 needed replacement or further analysis due to its low I.F. and I.D (see Appendix D). The rest of the items yielded difficulty according to the expected levels claimed by each quadrant of the assessment grid. In this test, 50% of items (6 different) have at least one distractor that needed revision (see Appendix E).

Table 7 shows the consolidated results of Test Two within the assessment grid. This time 61% of students correctly answered items in QII. Regarding these students, performance in QIII (39%) showed that they had little difficulty answering questions at low content/language demands and a little more difficulty with questions in QI (35%). Although, there was not a valid item to compare the levels of language difficulty in QIV, it seems that this group of students require more language support in order to perform better at the content demands, as they were able to answer items in both QIII and QI with a similar level of language demands but different demands in terms of content.

Table 7. Overall results, Test Two assessment grid

Q*	II	III						I		
Item	4	2	3	6	7	9	1	5	8	11
#**	54	45	37	29	22	41	40	32	20	30
%***	61	39						35		
Note:	* Quadrant in the assessment grid. ** Number of students who answered correctly each item. Numbers in the other quadrants are taken from the set of students who answered correctly item in QII. *** n=89									

Test Three

The last test, Test Three was applied as an achievement measure of the second term. In this case, 115 students took the tests in the same five groups. This test was developed taking into account the information shown in Table 8. The content components were defined according to the school curriculum. The language components were identified by the EFL teacher taking into account the curriculum, and the textbook. This time questions clearly differentiated whether students understood what adaptations are and how to explain them, or whether they had difficulties with the language used in understanding the questions.

Items in the assessment grid (Table 9) were carefully assigned to each quadrant as a result of the need to examine the item performance in terms of their relationship among each quadrant to spot the difference between language and content demands. Hence 50% of items had con-

textualized clues and the other 50% required students to recall concepts or understand without any support.

Table 8. Content and language components, Test Three

Topic	Living things
Components	**Scientific knowledge application** Use available information to support answers. **Inquire** Explain adaptations of living things according to their environment. **Explain phenomena** Identify adaptations in living things based on the characteristics of the ecosystem where they live. Based on the national standards released by the Colombian Ministry of Education and the school curriculum.
Language functions	Describing – Comparing – Observing – Predicting – Explaining
Language Structures	**Infinitive verbs:** Help sth to... to adapt, **Modal would:** would probably grow... **Present simple:** How does...? ... helps... **Relative clause pronouns:** that
Vocabulary	**Body Parts and adjectives:** thick feathers, huge lungs, long arms and tails, sharp teeth, waxy covering, warning colors, fins, wings, etc. **Adaptations:** migration, behavior, camouflage, morphological **Food Chain:** prey, predator **Habitats:** Ocean, desert, forest, mountains, South Pole, **Animals:** penguin, polar bear, frog, turtles, wolves, sharks, etc. **Verbs:** find food, find shelter, adapt, survive, travel, protect, escape, etc.

Table 9. Assessment grid, Test Three

<table>
<tr><td rowspan="2">Content Demands - Knowledge structure</td><td>High</td><td>Principles/relationships
Relationship between concepts – principles- processes - routines</td><td>Quadrant I
Defining: Item: 7
Identifying Items: 5 – 6 – 11</td><td>Quadrant II
Explaining Items: 9 – 12</td></tr>
<tr><td>Low</td><td>Concepts/classification
What? – Where? – Who? - When?</td><td>Quadrant III
Defining Item: 1
Identifying Items: 4 – 8 – 10</td><td>Quadrant IV
Explaining Items: 2 – 3</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>Lower-order Thinking skills/ CALP functions</td><td>Higher-order Thinking skills/ CALP functions</td></tr>
<tr><td colspan="2">Language Demands</td></tr>
</table>

Particularly, the assessment grid of Test Three showed a higher level of correspondence among items. This means, a question has at a minimum another question that measures similar knowledge or skills placed in another quadrant with a different level of demand. For instance, item 1 (QIII) the task posed to the learner was to define what adaptation is, parallels item 7 (QI) that aims at assessing whether the students know what adaptation is by comprehending its concept from a short text. The former item limits its language input to the question and the simple-statements of its answer options. The latter one demands a similar task but it includes reading the text and discarding other concepts from the options. Items 4, 8 and 10 (QIII) similarly correspond to items 5, 6, and 11 in QI. Likewise, items 2 and 3 QIV in comparison to Items 9 and 12 in QII.

The previous patterns of test design are relevant for the study because they allow for the examination of the role of the assessment grid for test development; whether it helped to discriminate between content and language demands of test items, or it did not. Hence, the item analysis, that follows, uncovered this concern and checked the reliability of each item.

Test Three had 12 items placed in each of the quadrants as follows: items 5, 7, and 11 in QI, items 9 and 12 in II, Items 1, 4, 8, and 10 in QIII, and Items 2 and 3 in QIV. A total of 3 items (2, 6 and 8) were found invalid, requiring further analysis or replacement (See Appendix I). This test had the fewer number of distractors to be revised in comparison to previous tests (see Appendix D).

Table 10 consolidates the results of Test Three. It is evident that students who answered correctly items in QII are better discriminated by the other quadrants. In detail, results show that students had a similar performance when language demands are minimum and the content demands vary. Performance in item 12 QIV (50%) revealed that students have better results when the language is more demanding (QIV 35%) than the content. A similar pattern is visible with item 9 in QII. 52% of the students had better performance (41%) at QIV in comparison to QIII (33%) and QI (35%).

Three tests were analyzed in order to identify their characteristics in terms of language and content demands, and placed their items in the assessment grid with the intention to discriminate which of the two constructs required more instruction, or have been mastered by students. By

Table 10. Overall results, Test Three assessment grid

Q*	II	III			I			IV
Item	12	1	4	10	5	7	11	3
#**	57	41	30	27	33	34	46	40
%***	50	29			33			35
Item	9	1	2	10	5	7	11	3
#**	60	38	46	30	36	40	53	47
%***	52	33			37			41
Note:	* Quadrant in the assessment grid. ** Number of students who answered correctly each item. Numbers in the other quadrants are taken from the set of students who answered correctly item in QII. *** n=115							

and large, it is evident that the assessment grid provides a valid framework to place the items. This information enriches the report of the tests by pointing out students' achievement by the levels of difficulty framed by each quadrant.

In general, tests items were mainly placed in QI and QII, suggesting that there is a high emphasis in assessment of content knowledge with low demand on language. Only Test Three had a valid item in QIV. This brings attention to the difficulty that may entail the design of low content demand questions with high language demands. The number of items that need revision varied from 1 to 3. A positive improvement was observed in the number of distractors that needed replacement. The assessment yielded useful categorization of items, in particular when they were related to each other in terms of content components.

DISCUSSION

There are two main contributions of this study. Firstly, it attempts to describe the summative assessment process that was actually carried out in a CLIL classroom, picturing the state of this curricular aspect from the inside. Although there is a lot of research on alternative assessment approaches (Short, 1993) aimed at obtaining accurate information about students' learning processes in formal education, summative tests, in their multiple-choice version, are still widely used to make decisions about students' promotion,

students' achievement, teacher performance, and even effectiveness of programs (Short, 1993). This study is evidence of this practice and how it is still rooted in classroom assessment yet at new curricular development approaches such as CLIL.

Sometimes assessment practices are flawed by practicability as the main way to judge tests. Elements such as validity and washback are vaguely applied. This study encourages the careful examination of tests, given its value aforementioned. So, teachers can evaluate their common assumptions by testing them systematically once in a while to guide their practice and enlighten their work with less subjectivity. An item analysis is a simple yet helpful instrument to build a set of informed decisions in test development.

Consequently, accepting that multiple-choice tests are pivotal in school dynamics, this study proposes an alternative to enriching this practice by using an assessment grid that reports distinctly students' achievement in terms of content and language demands. One of the most critical aspects in CLIL implementation is to establish this difference. According to test reports, generally the use of the assessment grid provides a valid framework to place test items in four different quadrants that combine the possible alternatives among content knowledge, thinking skills, and the required language to understand this at two levels of difficulty.

It is essential, though, to clarify that the assessment grid must be supported by a clear definition of the content and language components of each test, a consistent criterion to describe test items, and a valid set of items distributed in each of the quadrants. Besides, agreement on the levels of difficulty depends on the curricular outcomes suggested for the grade, in the case of the study, third grade.

In conclusion, the assessment grid allows reporting in detail the difficulties and the strengths of students after instruction or before it. This information could be helpful for CLIL teachers to increase their understanding of the language demands of any test item, to address specific strategies to actually attend students' needs, and afford foreign language learning beyond incidental language gains.

REFERENCES

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Pacific Grove, CA: Heinle & Heinle.

Congreso de Colombia. (1994). Ley 115. Ley General de Educación. Bogotá, Colombia: República de Colombia.

Congreso de Colombia. (2013). Ley 1651. Ley de Bilinguismo. Bogotá, Colombia: República de Colombia.

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

Curtis, A. (2012a). Colombian teachers' questions about CLIL: Hearing their voices – in spite of "the mess" (Part I). *Latin American Journal of Content and Language Integrated Learning*, *5*(1), 1–8. http://dx.doi.org/10.5294/laclil.2012.5.1.1

Curtis, A. (2012b). Colombian teachers' questions about CLIL: What can teachers' questions tell us? (Part II). *Latin American Journal of Content and Language Integrated Learning*, *5*(2), 1–12. http://dx.doi.org/10.5294/laclil.2012.5.2.6

Fandiño-Parra, Y. J., Bermúdez-Jiménez, J. R., & Lugo-Vásquez, V. E. (2012). The challenges facing the National Program for Bilingualism, Bilingual Colombia. *Educación Y Educadores*, *15*(3), 363–381. http://dx.doi.org/10.5294/edu.2012.15.3.2

Lo, Y. Y., & Lin, A. (2014). Designing assessment tasks with language awareness: Balancing cognitive and linguistic demands. *Assessment and Learning*, *3*, 97–119. Retrieved from http://wlts.edb.hkedcity.net/filemanager/file/A&L3(11)_Lo&Lin.pdf

Mariño, C. M. (2014). Towards implementing CLIL (content and language integrated learning) at CBS (Tunja, Colombia). *Colombian Applied Linguistics Journal*, *16*(2), 151. http://dx.doi.org/10.14483/udistrital.jour.calj.2014.2.a02

McDougald, J. S. (2009). The state of language and content instruction in Colombia. *Latin American Journal of Content and Language Integrated Learning*, *2*(2), 44–48. http://dx.doi.org/10.5294/laclil.2009.2.2.15

McDougald, J. S. (2015). Teachers' attitudes, perceptions and experiences in CLIL: A look at content and language. *Colombian Applied Linguistics Journal*, *17*(1), 25. http://dx.doi.org/10.14483/udistrital.jour.calj.2015.1.a02

Ministerio de Educación Nacional. (2004). *Programa nacional de bilingüismo*. Retrieved from Ministerio de Educación Nacional: http://www.mineducacion.gov.co/1621/articles-132560_recurso_pdf_programa_nacional_bilinguismo.pdf

Ministerio de Educación Nacional. (2006). *Estándares básicos de competencias en lenguas extranjeras: Inglés: Formar en lenguas extranjeras: ¡El reto! Lo que necesitamos saber y saber hacer*. Bogotá, Colombia: Imprenta Nacional. Retrieved from http://www.mineducacion.gov.co/cvn/1665/articles-115174_archivo_pdf.pdf

Ministerio de Educación Nacional. (2009). *Listado de colegios bilingües*. Retrieved from Colombia Aprende: http://www.colombiaaprende.edu.co/html/home/1592/article-228186.html

Ministerio de Educación Nacional. (2013). *Programa nacional de inglés 2015-2025*. Retrieved from Colombia aprende: http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-343287_recurso_1.pdf

Otálora, B. (2009). CLIL research at Universidad de La Sabana in Colombia. *Latin American Journal of Content and Language Integrated Learning*, *2*(1), 46–50. http://dx.doi.org/10.5294/laclil.2009.2.1.7

Pacific Policy Research Center. (2010). *Successful bilingual and immersion education models/programs*. Retrieved from Kamehameha Schools: http://www.ksbe.edu/_assets/spi/pdfs/Bilingual_Immersion_full.pdf

Quartapelle, F. (Ed.). (2012). *Assessment and evaluation in CLIL*. Pavia, Italy: Ibis. Retrieved from http://aeclil.altervista.org/Sito/wp-content/uploads/2013/02/AECLIL-Assessment-and-evaluation-in-CLIL.pdf

Rodriguez, M. (2011). CLILL: Colombian Leading into Content Language Learning. *Íkala, Revista de Lenguaje y Cultura 16(2),* 79-89. Retrieved from https://aprendeenlinea.udea.edu.co/revistas/index.php/ikala/article/view/9912

Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly*, *27*(4), 627. http://dx.doi.org/10.2307/3587399

APPENDIX A

Test One

Table 11 shows the analysis of Test One using Wesche's (1983) "Components of a Test".

Table 11. Analyzing Test One with Wesche's (1983) "Components of a Test"

	Wesche's Components of Test One			
Test Item	**Stimulus Material**	**Task posed to the learners**	**Learners' Response**	**Scoring Criteria**
1	a) Direct Task statement b) Options in pictures.	Identify among the groups a group of living things (What).	Choose the correct group of living things	Correct / incorrect answer.
2	a) Contextualized Task statement. b) Table with simple statements. c) Options in pictures.	Identify characteristics of a living thing.	Choose the correct living thing.	
3	a) Task statement. b) Table. c) Direct question. d) Options in simple statements.	Relate concepts to the table and identify processes.	Choose the best description of a table.	
4	a) Compound statement b) Table. c) Question. d) Options in simple statements (Academic vocabulary)	Identify the relationship between the concept of biotic factors and the examples.	Choose the description of the table.	
5	a) Task statement. b) Picture. c) Task statement d) Options in pictures.	Identify abiotic factor necessary for any living thing to grow. (What)	Choose the correct abiotic factor.	
6	a) Task statement. b) Complete a statement c) Single-word options.	Identify abiotic factors. (What)	Choose the word that fills well the blank.	
7	a) Compound statement. b) Picture. c) Task statement. d) Options in simple statements.	Identify the relationship among abiotic factors and biotic factors in the experiment.	Choose the best description of the picture.	
8	a) Task statement. b) Picture. c) Task statement d) Options in pictures.	Identify the cause of an event.	Choose the picture that explains the problem.	

	Wesche's Components of Test One			
Test Item	**Stimulus Material**	**Task posed to the learners**	**Learners' Response**	**Scoring Criteria**
9	a) Task statement. b) Picture c) Single-word options. (Academic vocabulary)	Identify the domain (What).	Choose the correct domain.	Correct / incorrect answer.
10	a) Task statement. b) Table. c) Single-word options. (Academic vocabulary)	Identify kingdoms and domains among the pictures of the table.	Choose the correct heading.	
11	a) Task statement. b) Pictures. c) Options in complex sentences. (Academic vocabulary)	Explain concepts.	Choose the best explanation.	
12	a) Task statement. b) Table c) Question. d) Options in simple statements. (Academic vocabulary)	Identify the relationship among concepts and examples.	Choose the best description.	

Table 12 shows item facility and item discriminability for Test One.

Table 12. Item facility & item discriminability (n=89), Test One

Item	# correct answers	I.F.	I.D.
1	53	0.60	0.49
*2	74	0.83	0.44
3	58	0.65	0.44
*4	16	0.18	0.28
5	39	0.44	0.45
6	35	0.39	0.46
7	51	0.57	0.47
8	54	0.61	0.5
9	49	0.55	0.23
10	33	0.37	0.48
11	24	0.27	0.19
12	40	0.45	0.56

Table 13 shows the distractor analysis for Test One.

Table 13. Distractor analysis (n=89), Test One

Item	A	B	C	D	W	Z	A	B	C	D	W	Z
1	7	*53	6	17	5	1	8	60	7	19	6	1
2	4	7	*74	3	1	0	4	8	83	3	1	0
3	9	9	*58	11	0	2	10	10	65	12	0	2
4	19	*16	12	37	0	5	21	18	13	42	0	6
5	*39	2	41	4	0	3	44	2	46	4	0	3
6	24	*35	19	1	0	10	27	39	21	1	0	11
7	*51	9	13	8	0	8	57	10	15	9	0	9
8	16	2	*54	7	0	10	18	2	61	8	0	11
9	*49	20	11	8	0	1	55	22	12	9	0	1
10	9	*33	18	27	0	2	10	37	20	30	0	2
11	25	*24	29	8	0	3	28	27	33	9	0	3
12	24	13	*40	9	0	1	27	15	45	10	0	1

Test Two

Table 14 shows the item facility and item discriminability for Test Two.

Table 14. Item Facility & Item Discrimination (n=89), Test Two

Item	#	I.F.	I.D.
1	63	0.71	0.4
*2	78	0.88	*0.2
3	54	0.61	0.32
4	54	0.61	0.34
5	47	0.53	0.47
6	36	0.40	0.59
7	43	0.48	0.45
8	35	0.39	0.51
9	64	0.72	0.45
10	27	0.30	0.31
11	54	0.61	0.45
12	55	0.62	0.44

Table 15 shows a distractor analysis for Test Two.

Table 15. Distractor analysis (n-89), Test Two

Item	A	B	C	D	A	B	C	D
1	*63	5	3	17	71	6	3	19
2	3	*78	5	3	3	88	6	3
3	4	20	*54	10	4	22	61	11
4	12	7	14	*54	13	8	16	61
5	*47	10	1	30	53	11	1	34
6	7	*36	30	16	8	40	34	18
7	13	28	*43	3	15	31	48	3
8	12	13	28	*35	13	15	31	39
9	*64	16	5	4	72	18	6	4
10	39	*27	10	12	44	30	11	13
11	5	14	*54	14	6	16	61	16
12	10	9	13	*55	11	10	15	62

Test Three

Table 16 shows the item facility and item discrimination for Test Three.

Table 16. Item facility & item discrimination (n=115), Test Three

Item	#	I.F.	I.D.
1	88	0.77	0.4
2	34	0.30	0.3
3	80	0.70	0.4
4	55	0.48	0.4
5	58	0.50	0.5
6	34	0.30	0.3
7	74	0.64	0.2
8	103	0.90	0.3
9	60	0.52	0.5
10	47	0.41	0.4
11	92	0.80	0.4
12	57	0.50	0.4

Table 17 shows the distractor analysis for Test Three.

Table 17. Distractor analysis (n=115), Test 3

Item	A	B	C	D	A	B	C	D
1	*88	8	13	6	77	7	11	5
2	23	*34	37	21	20	30	32	18
3	8	18	*80	9	7	16	70	8
4	12	17	30	*55	10	15	26	48
5	13	*58	13	29	11	50	11	25
6	*34	32	22	26	30	28	19	23
7	9	14	*74	18	8	12	64	16
8	4	7	0	*103	3	6	0	90
9	40	7	*60	6	35	6	52	5
10	19	*47	31	14	17	41	27	12
11	8	9	*92	5	7	8	80	4
12	21	23	*57	13	18	20	50	11